Comparing building blocks of life

Pirovano, W.A.

2010

document version
Publisher's PDF, also known as Version of record

VRIJE UNIVERSITEIT

Comparing building blocks of life: sequence alignment and evaluation of predicted structural and functional features

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. L.M. Bouter,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Exacte Wetenschappen
op vrijdag 15 januari 2010 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Walter Alexander Pirovano

geboren te Gouda

promotor:	prof.dr. J. Heringa
copromotor:	dr.ir. K.A. Feenstra

Contents

Preface

Alignment, beer and coffee.

Somewhere half-way my Master project, the bioinformatics lab of Professor Pesole had the great honour of receiving the alignment guru Des Higgins. For a few months now I had been struggling with my multiple alignments and I could not imagine a presentation title more promising than 'Everything you ever wanted to know about multiple alignments but were too shy to ask'. Actually the whole group was looking forward to the talk of 'Mr. ClustalW (Mister Clustal double-U)'.

To be honest, I do not remember a lot of that presentation, though one slide in particular has made a big impression on me: a nice picture of a typical Irish pub where, in between one beer and another, the whole Clustal-idea was conceived. That sounded pretty much inspiring: combining beer and sequence alignments. I therefore decided not to bother him with questions about algorithmic details, but rather ask him if he'd knew someone in the Netherlands where these two fields could be combined. He suggested to ask Jaap Heringa for a PhD position: 'you will certainly have a great time with him'. Des told me he had enjoyed very much working together with Jaap on a brand new multiple alignment algorithm called T-Coffee. If eyes could talk, mine would have certainly said: 'alignment, beer and coffee', that's over the top!

Sidetracks.

A quick internet survey 'Jaap Heringa' brought me to London, whereas Valeria and I had set our minds on Holland, or better Amsterdam (which are two different things). I forgot about the coffee-story and tried to search for alternatives. It occurred though that in Nijmegen and Groningen there were interesting labs and I wrote them open applications for a PhD position. Very kind invitations followed and in February

2005 I flew to the Netherlands for my very first job interviews. In Nijmegen I had a nice conversation with Professor Siezen. At some point we were talking about bioinformatics in The Netherlands in general when he asked:

It so happened that ...

'will you also visit the bioinformatics lab of Jaap Heringa at the VU?' 'Well, actually, no, but isn't he in the UK?' 'No, he moved to Amsterdam quite a while ago. Just try to phone him and see what he is up to ...'. Fine, great I thought, but will it work out if I only have one morning left before I have to return to Milan? But who said 'There's no such thing as luck'? Well, I think there is: 1) Jaap immediately answered his phone and 2) yes, I could come and visit the lab the next morning and give a presentation of my thesis work during their weekly groupmeeting. And so it happened that three weeks after my visit I received the following mail:

Hi Walter,

Deze week bereikte ons het bericht dat het Siezen/Heringa voorstel voor Bio-Range is geaccepteerd door het Netherlands Bioinformatics Centre (NBIC). Ik heb inmiddels met Prof. Roland Siezen overlegd en zou je daarom willen vragen of je nog steeds zin hebt om naar Amsterdam te komen om je Ph.D te doen.

Wanneer dat zo is, laat dan ook even weten wanneer je zou kunnen/willen beginnen.

Prettige Paasdagen,
Hartelijke groet,

Jaap

I still remember the happiness of my mother and twinkling eyes of Valeria.

This is where the journey started ...

CHAPTER 1

Introduction

1.1 Life in the (post-)genomic era

1.1.1 The Human Genome Project (HGP)

One of the most intruiging and widely-used terms used in biology is 'the post-genomic era'. It has no unequivocal definition, but many scientists will agree that it refers to the period after the year 2000 in which a first draft of the human genome became available (Lander et al., 2001; Venter et al., 2001). The initiative to 'decode the human book of life' took off in 1990 and was carried out by major institutes and universities around the globe working together on the Human Genome Project (HGP).

At that time several complete genome sequences from other organisms were already available. An important pioneer in this field was Fred Sanger who determined the first genomic DNA sequence in 1977 (Sanger et al., 1977): the 5368 base-pair genome of the Φ-X174 virus. Subsequent development of new sequencing techniques resulted in a rapid growth of the number of available genomes over the years. This period is often referred to as the 'genomic era' where the main objective was to physically determine the sequential order of the building blocks of genomes: the nucleotides A (Adenine), C (Cytosine), T (Thymine) and G (Guanine). The human genome however constituted the ultimate challenge because of the large number of nucleotides, over 3 billion, that had to be cataloged. At the time the HGP started it was commonly believed that by unraveling our own DNA all kinds of biological processes and genetic diseases would become much better understood. The HGP project certainly was a very challenging and fundamental piece of work. Using the words of the former US President Bill Clinton, it was without a doubt 'a milestone for humanity'. In the

end the HGP itself was a great success, although the project did not only bring good news ...

1.1.2 'Good news and bad news ...'

The good news was that the project turned out to be much less time and money consuming than had been anticipated at the start, even though it was not totally clear at the time which sequencing techniques should have been applied. We all know that prestigious engineering projects usually show the opposite behaviour (Dutch examples include the construction of the HSL high-speed railway and the Amsterdam Noord-Zuid lijn metro). It is therefore an extraordinary result that within only 10 years time and for less than 3 billion dollars scientists succeeded in elucidating a similar amount of base-pairs that constitute their genetic code. The bad news was, and still is, that 'sequencing' a genome appeared to be a very different task than 'deciphering' it. In fact the genetic code itself constitutes only a tiny layer of the complexity of biological systems. Even now that we have unmasked every piece of human DNA, we have only very little clue about how the 20.000 genes are regulated and how they cooperate. Also at a higher level we have to admit that mechanisms underlying cellular processes at network levels often remain obscure. At this point we have silently entered the so-called 'post-genomic' era. The major focus of the community has now been shifted to deciphering the functions that are encoded in the nucleic acids rather than cracking their sequential order.

1.1.3 The Golden Age of Bioinformatics

The world famous Dutch football-player Johan Cruyff once said: 'every disadvantage has its advantage'. Bioinformaticians could now take advantage of the situation and started tackling the huge list of unsolved questions that arose from the human genome and subsequent other genome projects. Although the term 'bioinformatics' had already been coined in 1978 by the Dutch theoretical biologist Paulien Hogeweg, the actual bioinformatics claim to fame was only grasped after 2000 when the importance of this field was also 'politically' discovered. This has led to a tremendous increase in popularity of bioinformatics in general and has also paved the road for many 'omics'-related research areas that study the interactions of biological data at different levels (such as the genome, transcriptome, proteome and metabolome). At the same time also the field of systems biology is rapidly evolving and seeks to understand complex biological systems or processes as a whole rather than considering each single constituent. It should also be stressed that over time bioinformatics itself has changed as well. In the early days, algorithmic development of analysis tools was mainly hypothesis driven due to a lack of data, whereas nowadays we use large-scale data analysis to discover biological trends.

Over the past years many algorithms and analysis tools have been developed to get a better grip on the overwhelming amount of sequence data available. An important collaborative scientific initiative has been the Encode Project (ENCODE Project

Consortium et al., 2007) that aimed to functionally annotate 1% of the human genome through the cooperative effort of many research institutes from all over the world. The project has led to many new insights, such as the apparent importance of microRNAs in protein regulation. However, only little is understood about these microRNAs and (cell) regulatory processes in general. A perhaps even more striking example is given by the growing difficulties in managing and analyzing Next-Generation Sequencing (NGS) data. NGS technologies have revolutionarised genome sequencing projects as they have sped up the process over a million fold. Nonetheless, the rapidly increasing volumes of sequence data coming out of the NGS machines is difficult to handle and to mine. The creation of new solutions for more efficient data storage and better pipelines for subsequent analysis is a challenge that can only be tackled by the joint effort of the bioinformatics community in close collaboration with biologists and informaticians. As a result it is likely that this century will keep bioinformaticians rather busy.

1.2 Molecular evolution

The overwhelming diversity of present-day DNA, RNA and protein sequences is the result of processes of molecular evolution. The fundamental ingredients that are required to get Darwinian evolution include a template (the DNA sequence), a copying or reproduction mechanism, sequence variation and selection. Along the evolutionary road, sequence templates have undergone a large number of variations we call 'sequence edits'. These can be grouped into several categories:

- substitutions: the modification of one nucleotide into another;
- insertions and deletions: the addition/removal of nucleotides into/from the DNA; these modifications may also comprise gains or losses of chromosomal fragments;
- transpositions: mobile stretches of DNA that can move from one genomic position to another;
- recombination: shuffling of genes during reproduction events.

As a consequence, sequence edits induce genetic diversity which provides the playground for 'natural selection'. In this scenario some variations or modifications are preserved because of the structural or functional benefits for the organism. The majority of evolutionary changes are subject to 'neutral selection' rather than by selective pressure. The neutral theory of molecular evolution, originally proposed by Kimura (1979), implies that most sequence edits occur randomly and do not affect the fitness of an organism. Subsequently the apparent 'harmlessness' of most modifications leads to a so-called random 'genetic drift'.

It would be fascinating to watch a movie that shows us step by step the mechanisms and changes that have resulted in the world's astonishing rich flora and fauna.

Unfortunately that movie does not exist. Nowadays we only have very limited knowledge about ancestral species and can only simulate evolutionary principles. Evolution proceeds in two major modes: divergent and convergent evolution. The most common type in nature is divergent Darwinian evolution, which implies that existing species give rise to new species and that similarity between their characters can be explained by their common ancestry. Divergent evolution of biological sequences is the result of speciation events as well and leads to similar sequences in different species that share comparable functions. This brings us to one of the key concepts in (molecular) biology: homology. Two sequences are defined to be homologous if they share a common ancestor. In contrast, convergent evolution implies that similarity between species characters or sequences is the results of independent parallel evolutionary processes. Examples include the wings of bats and birds, a characteristic which was not present in their common ancestor, the independent development of the eyes in vertebrates and cephalopods (*e.g.* squids), or two originally unrelated proteins that over evolutionary time adopt similar enzymatic functions due to some (random) mutations in the active site (eg chymotrypsin and subtilisine).

1.3 Homology: a sparring partner of bioinformatics

Homology itself can be further split into two categories: orthology and paralogy. Orthologous relationships between sequences are due to a speciation event that leads to two genes in different organisms that carry out similar functions. Paralogous relationships instead originate after a gene duplication event that leads to two related genes that may carry out different functions. Especially the detection of orthologous relationships provides clues for function elucidation and might answer central questions like: what does this gene do or will this protein bind to a ligand?

To fully understand this concept, it is important to stress that genes (or proteins) that have arisen from a common ancestor do not only share similar functions, but also share similarities between their sequences. Moreover, the most conserved sequence parts indicate regions that are often essential for a proper functioning of the gene. Other regions might show more variability due to the accumulation of mutation events, though not necessarily alter the general function. In other words, the function of a gene or protein is more conserved than the physical sequence itself. Different rules however apply to cases of convergent evolution where a general lack in sequence overlap can be attributed to the absence of a common origin.

A significant example of the practical application of homology could be the following. Suppose we have a human gene A of which we suspect that it is a key player in a certain type of cancer. We may compare its DNA sequence to the mouse genome and eventually discover a similar piece of DNA in mice that would indicate the existence of a mouse gene A counterpart. Subsequent experiments in mice could provide more evidence for the importance of gene A in cancer. In other words, given that the function of a sequence is more conserved than the sequence itself, we have a solid basis for the transfer of function through homology.

It should be kept in mind though that function, structure and sequence comparisons follow different metrics and that in principle these can not be directly compared to each other. For instance, a high sequence identity shared between two proteins does not imply a small root mean square deviation (RMSD) between their structures or shared ontologies from a functional point of view. Especially in this field the concept of 'one-size-fits-all' is simply not applicable.

1.4 Sequence analysis

Given the ever increasing amount of sequencing projects, the amount of sequence data is growing rapidly. To give an impression, currently genomes are available for over 5000 species covering all three domains of life and viruses. Meanwhile also the gap with functionally annotated data is increasing and the introduction of new high-throughput techniques (commonly referred to as Next-Generation Sequencing), such as the Illumina sequencing technology and the Roche 454 sequencing system, will further widen the gap. In terms of cost, we are getting closer to the '1000 dollar genome' ($1000 for a fully sequenced human genome) (Service, 2006) though its full analysis has an inestimable price-tag. Concerning proteins, the large 'non-redundant' (nr) protein database contains millions of sequences though the amount of experimentally determined structures and functions is significantly smaller (*i.e.* to date the PDB structure database contains around 60.000 solved crystal structures).

A successful attempt to somehow narrow this gap is made by the bioinformatics discipline 'sequence analysis'. As the name partly implies, its main focus is on assigning biological functions to unannotated DNA, RNA and protein sequence data. This goal is mainly achieved by performing comparative analyses between 'query' sequences and sequences with known functions (for a review see Heringa and Pirovano, 2007). In this manner functional information can be transferred from one sequence to another as described in the previous section. Probably the best example of such a method is the BLAST program (Altschul et al., 1990), which is one of the most popular and widely used tools among bioresearchers. Provided with a query sequence, the method is able to scan huge (annotated) sequence databases for sequence similarity in only a limited amount of time. The huge success of sequence analysis tools like BLAST can be explained by the fact that they can help biologists in guiding (or even replacing) expensive experiments. Main benefits are not only reduction in costs, but also in terms of time as computer analyses are much faster than full experiments. On the other hand, the value of comparative computer analysis is largely dependent on the accuracy of experimental annotations. Wet-lab and computer research should therefore reciprocally benefit from each other in a synergistic co-existence. A good example here is the integration of interaction and kinetics data, models of biological processes and simulated network models, which together provide a context for data analysis and annotation. These type of cell context approaches are widely used in the field of systems biology and are expected to gain more and more importance in bioinformatics as well.

1.5 (Multiple) sequence alignment

A major problem arising in comparative analyses concerns the fact that sequences, *e.g.* two homologous proteins occurring in different species, usually have dissimilar lengths and compositions. These differences are the tangible result of evolutionary processes during which some (or a lot of) amino acids within a protein have been changed into others, were inserted into or disappeared from a sequence (so-called indels). A very useful answer to this problem is given by 'sequence alignment'. A more formal definition would state that the goal of sequence alignment is to align a set of input sequences (either DNA, RNA or protein) in such way that their evolutionary relationships are best represented. From a more practical point of view alignment seeks to arrange either the nucleic or amino acids blocks in such a way that the number of identical or similar blocks is maximized throughout the columns. This is obtained by introducing whitespaces, called 'gaps', within the sequences thus conferring to each sequence the same length. Sequence alignment nowadays is a fundamental starting point for many other types of analyses, among which structure and function prediction, evolutionary analysis and motif detection.

The term 'pairwise alignment' implies the comparison of solely two input sequences. Some powerful algorithms have been developed years ago to find the optimal arrangement of the blocks (Needleman and Wunsch, 1970; Smith and Waterman, 1981) and lie for instance at the basis of BLAST database searches. Multiple sequence alignment is clearly used to compare three or more biological sequences. An important issue here though is that the complexity in finding an optimal arrangement increases exponentially with the number of sequences added to the alignment. In practice, it is unfeasabile to find the optimal solution for more than ten sequences due to the enormous amount of letter combinations that has to be explored. To somehow overcome speed-related problems an frequently used concept is 'heuristics', which means that, according to certain criteria, shortcuts are taken to reduce the search space. As a consequence heuristics can help to find a near-optimal solution in a much shorter time span but without any guarantees as to the optimality of the solution found. In more recent years there have been important developments in this direction leading to fast tools which are very well appreciated in the community. To stress the importance of heuristic methods, it is worthwhile noticing that both BLAST and the multiple alignment tool ClustalW (Thompson et al., 1994) have been cited over 25.000 times since their appearance in the 90's (source: `www.isiknowledge.com`).

The exciting area of multiple sequence alignment is evolving rapidly, which is underlined by the large amount of methods and strategies that have been introduced over the past few years. In this thesis important improvements of our widely-used protein multiple alignment tool PRALINE are presented and we feel that the new insights make a significant contribution to the field (Pirovano et al., 2008b, 2009a). Furthermore we have performed an extensive study on the interpretation of multiple sequence alignment regarding the functional analysis of protein subfamilies from multiple sequence alignments (Pirovano et al., 2006; Feenstra et al., 2007). In summary, this study has been an attempt to make a contribution both at an algorithmic and

an analytical/practical level.

1.6 Thesis outline

As stated in the previous section, the central theme of this thesis is protein multiple sequence alignment. After an extended introduction on the topic, we will treat this issue from three different perspectives:

1. Can we improve current multiple alignment protocols by including predicted structural knowledge?
2. In which way can we detect functional specificity entailed in alignments?
3. What is the meaning of alignment in general and how should we interpret protein dynamics?

Chapter 2 gives a general overview of protein multiple sequence alignment. First it provides background information including early and recent breakthroughs. Then a description of state-of-the-art alignment methods and visualization tools is given. Finally practical protocols are provided to help researchers in building a reliable multiple sequence alignment (Pirovano and Heringa, 2008).

In Chapter 3 and 4 we describe new strategies for improving the existing multiple alignment program PRALINE by using information gained from predicted structural elements. Since a protein's structure is more conserved than its sequence, we have attempted to incorporate this knowledge to 'guide the alignment'. Chapter 3 has its main focus on improved alignment of transmembrane proteins, a special class of proteins that are characterized by their settlement in cellular membranes. We achieve this goal by accurately predicting the location of transmembrane segments and subsequent application of an alternative evolutionary scoring scheme tailored to these regions (Pirovano et al., 2008b). In Chapter 4 we extend this approach to a more general application which can be used for the whole protein spectrum. Here we use predicted secondary structure information in combination with homology-extended alignment to enhance the alignment quality (Pirovano et al., 2009a). In addition an advanced webserver toolbox is presented which allows a full combination of all strategies proposed.

In Chapter 5 and 6 we take a closer look at the functional information that is entailed in multiple sequence alignments. In particular we zoom in on the issue of functional specificity which can explain functional dissimilarity between protein subfamilies. Chapter 5 introduces the 'Sequence Harmony method' that provides a new entropy-based measure for detecting specificity (Pirovano et al., 2006). Whereas conventional methods link specificity determining alignment positions to conserved residue patterns within subfamilies, our algorithm captures subfamily differences without imposing sequence conservation. In addition it takes neighbouring residues into account to determine the intensity of a specificity signal. The performance is tested on experimentally verified mutation data and demonstrates that the method accurately

selects known functional sites. In Chapter 6 we present the Sequence Harmony webserver which offers a quick and intuitive analysis of specificity determining residues on the web (Feenstra et al., 2007). Moreover it allows to map the functional residues onto 3D structural data. The user is guided through all stages of the analysis by means of the biologically example of plant alternative oxidases.

In Chapter 7 and 8 we focus on the relationship between structure and sequence alignments. Again the central point is that related proteins show a higher degree of conservation between their structures compared to the sequence level. As a consequence the structural superposition of protein 3D structures can provide a fundamental basis in deriving principles of sequence relationships. For example the quality of sequence alignment routines is evaluated on gold standard alignments are derived from structural comparisons. However, in spite of the fact that protein structures are dynamic, structure and sequence alignments are often presented as static snapshots. The main goal of this study was to estimate the effects of structural diversity on both structure and derived sequence alignments. We observed that even small structural changes can lead to severe differences in the derived sequence alignments (Pirovano et al., 2008a). As a consequence there is no unique best alignment representation possible. In Chapter 8 we further explore these 'flexible alignment regions' by studying the structural features and functional importance they entail (Pirovano et al., 2009b).

Chapter 9 contains the summarizing discussion of this PhD work. The main findings described in this thesis are collected attempting to provide answers to the above stated research questions. In conclusion we elaborate the results and discuss some future directions in the area of multiple sequence alignment.

CHAPTER 2

Multiple sequence alignment

Published as:

Pirovano, W., and Heringa, J. (2008).
Multiple sequence alignment.
Methods Mol. Biol., 452:143–161.

Abstract

Multiple sequence alignment (MSA) has assumed a key role in comparative structure and function analysis of biological sequences. It often leads to fundamental biological insight into sequence-structure-function relationships of nucleotide or protein sequence families. Significant advances have been achieved in this field, and many useful tools have been developed for constructing alignments. It should be stressed, however, that many complex biological and methodological issues are still open. This chapter first provides some background information and considerations associated with MSA techniques, concentrating on the alignment of protein sequences. Then, a practical overview of currently available methods and a description of their specific advantages and limitations are given, so that this chapter might constitute a helpful guide or starting point for researchers who aim to construct a reliable MSA.

2.1 Introduction

2.1.1 Definition and implementation of an MSA

A multiple sequence alignment (MSA) involves three or more homologous nucleotide or amino acid sequences. An alignment of two sequences is normally referred to as a pairwise alignment. The alignment, whether multiple or pairwise, is obtained by inserting gaps into sequences such that the resulting sequences all have the same length L. Consequently, an alignment of N sequences can be arranged in a matrix of N rows and L columns, in a way that best represents the evolutionary relationships among the sequences.

Organizing sequence data in MSAs can be used to reveal conserved and variable sites within protein families. MSAs can provide essential information on their evolutionary and functional relationships. For this reason, MSAs have become an essential prerequisite for genomic analysis pipelines and many downstream computational modes of analysis of protein families such as homology modeling, secondary structure prediction, and phylogenetic reconstruction. They may further be used to derive profiles (Gribskov et al., 1987) or hidden Markov models (Haussler et al., 1993; Bucher et al., 1996) that can be used to scour databases for distantly related members of the family. As the enormous increase of biological sequence data has led to the requirement of large-scale sequence comparison of evolutionarily divergent sets of sequences, the performance and quality of MSA techniques is now more important than ever.

2.1.2 Reliability and evolutionary hypothesis

The automatic generation of an accurate MSA is computationally a tough problem. If we consider the alignment or matching of two or more protein sequences as a series of hypotheses of positional homology, it would obviously be desirable to have *a priori* knowledge about the evolutionary (and structural) relationships between the

sequences considered. Most multiple alignment methods attempt to infer and exploit a notion of such phylogenetic relationships, but they are limited in this regard by the lack of ancestral sequences. Naturally, only observed taxonomic units (OTUs), *i.e.*, present-day sequences, are available. Moreover, when evolutionary distances between the sequences are large, adding to the complexity of the relationships among the homologous sequences, the consistency of the resulting MSA becomes more uncertain (see Note 1 in Section 2.4).

When two sequences are compared it is important to consider the evolutionary changes (or sequence edits) that have occurred for the one sequence to be transformed into the second. This is generally done by determining the minimum number of mutations that may have occurred during the evolution of the two sequences. For this purpose several amino acid exchange matrices, such as the PAM (Dayhoff et al., 1978) and BLOSUM (Henikoff and Henikoff, 1992) series, have been developed, which estimate evolutionary likelihoods of mutations and conservations of amino acids. The central problem of assembling an MSA is that a compromise must be found between the evolutionarily most likely pairwise alignments between the sequences, and the embedding of these alignments in a final MSA, where changes relative to the pairwise alignments are normally needed to globally optimize the evolutionary model and produce a consistent multiple alignment.

2.1.3 Dynamic programming

Pairwise alignment can be performed by the dynamic programming (DP) algorithm (Needleman and Wunsch, 1970). A two-dimensional matrix is constructed based on the lengths of the sequences to be aligned, in which each possible alignment is represented by a unique path through the matrix. Using a specific scoring scheme, which defines scores for residue matches, mismatches, and gaps, each position of the matrix is filled. The DP algorithm guarantees that, given a specific scoring scheme, the optimal alignment will be found. Although dynamic programming is an efficient way of aligning sequences, applying the technique to more than two sequences quickly becomes computationally unfeasible. This is due to the fact that the number of comparisons to be made increases exponentially with the number of sequences. Carrillo and Lipman (1988) and more recently Stoye et al. (1997) proposed heuristics to reduce the computational requirements of multidimensional dynamic programming techniques. Nonetheless, computation times required remain prohibitive for all but the smallest sequence sets.

2.1.4 The progressive alignment protocol

An important breakthrough in multiple sequence alignment has been the introduction of the progressive alignment protocol (Feng and Doolittle, 1987). The basic idea behind this protocol is the construction of an approximate phylogenetic tree for the query sequences and repeated use of the aforementioned pairwise alignment algorithm. The tree is usually constructed using the scores of all-against-all pairwise alignments

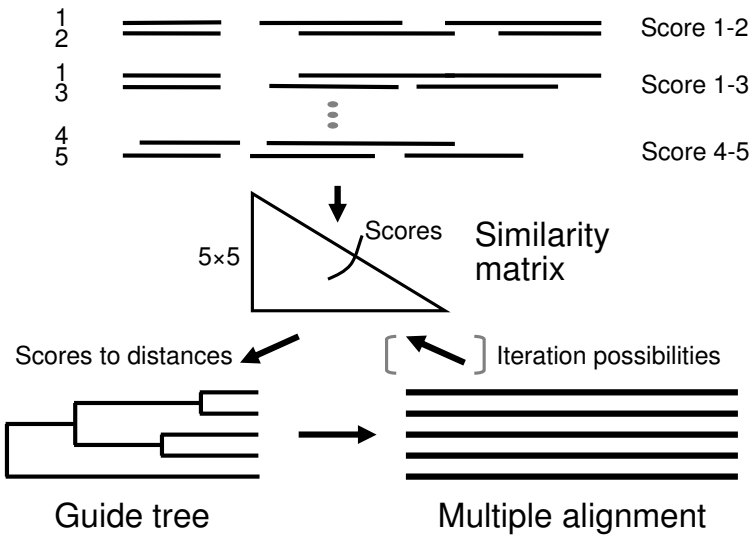


Figure 2.1: Schematic representation of the progressive alignment protocol. A similarity (distance) matrix, which contains scores from all pairwise alignments, is used to construct a guide tree. The final alignment is built up progressively following the order of the guide tree. The black arrow between brackets indicates possible iterative cycles.

across the query sequence set. Then the alignment is build up by progressively adding sequences in the order specified by the tree (see Figure 2.1), which is therefore referred to as the *guide tree*. In this way, phylogenetic information is incorporated to guide the alignment process, such that sequences and blocks of sequences become aligned successively to produce a final MSA. Fortunately, as the pairwise DP algorithm is only repeated a limited number of times, typically on the order of the square of the number of sequences or less, the progressive protocol allows the effective multiple alignment of large numbers of sequences.

However, the obtained accuracy of the final MSA suffers from the so-called greediness of the progressive alignment protocol; that is, alignment errors cannot be repaired anymore and will be propagated into following alignment steps ('Once a gap, always a gap'). In fact, it is only later during the alignment progression that more information from other sequences (*e.g.*, through profile representation) (Gribskov et al., 1987) becomes employed in the alignment steps.

2.1.5 Alignment iteration

Triggered by the main pitfall of the progressive alignment scenario, some methods try to alleviate the greediness of this strategy by implementing an iterative align-

ment procedure. Pioneered by Hogeweg and Hesper (1984), iterative techniques try to enhance the alignment quality by gleaning increased information from repeated alignment procedures, such that earlier alignments are 'corrected' (Hogeweg and Hesper, 1984; Gotoh, 1996). In this scenario, a previously generated MSA is used for improvement of parameter settings, so that the initial guide tree and consequently the alignment can be optimized. Apart from the guide tree, the alignment procedure itself can also be adapted based on observed features of a preceding MSA. The iterative procedure is terminated whenever a preset maximum number of iterations or convergence is reached. However, depending on the target function of an iterative procedure, it does not always reach convergence, so that a final MSA often depends on the number of iterations set by the user. The alignment scoring function used during progressive alignment can be different from the target function of the iteration process, so a decision has to be made whether the last alignment (with the maximal iterative target function value) or the highest scoring alignment encountered during iteration will be taken as the final result upon reaching convergence or termination of the iterations by the user.

Currently, a number of alternative methods are able to produce high-quality alignments. These are discussed in Section 2.3, as well as the options and solutions they offer, also with respect to the considerations outlined in the preceding.

2.2 Materials

2.2.1 Selection of sequences

Since sequence alignment techniques are based upon a model of divergent evolution, the input of a multiple alignment algorithm should be a set of homologous sequences. Sequences can be retrieved directly from protein sequence databases, but usually a set is created by employing a homology searching technique for a provided query sequence. Widely used programs such as BLAST (Altschul et al., 1990) or FASTA (Pearson, 1990) employ carefully crafted heuristics to perform a rapid search over sequence databases and recover putative homologues. Selected sequences should preferably be orthologous but in practice it is often difficult to ensure that this is the case. It is important to stress that MSA routines will also be capable of producing alignments of unrelated sequences that can appear to have some realistic patterns, but these will be biologically meaningless ('garbage in, garbage out'). For example, it is possible that some columns appear to be well conserved, although in reality no homology exists. Such misinterpretation could well have dramatic consequences for conclusions and further analysis modes. Although the development of P- and E-values to estimate the statistical significance of putative homologues found by homology searching techniques limits the chance of false positives, it is entirely possible that essentially non-homologous sequences enter the alignment set, which might confuse the alignment method used.

2.2.2 Unequal sequence lengths: global and local alignment

Query sequence sets comprise sequences with unequal length. The extent of such length differences requires a decision whether a *global* or *local* alignment should be performed. A *global* alignment strategy (Needleman and Wunsch, 1970) aligns sequences over their entire length. However, many biological sequences are modular and contain shuffled domains (Heringa and Taylor, 1997), which can render a global alignment of two complete sequences meaningless (see Note 2 in Section 2.4). Moreover, global alignment can also lead to incorrect alignment when large insertions of gaps are needed, for example, to match two domains A and B in a two-domain protein against the corresponding domains in a three-domain structure ACB. In general, the global alignment strategy is appropriate for sequences of high to medium sequence similarity. At lower sequence identities, the global alignment technique can still be useful provided there is confidence that the sequence set is largely colinear without shuffled sequence motifs or insertions of domains. Whenever such confidence is not present, the *local* alignment technique (Smith and Waterman, 1981) should be attempted. This technique selects and aligns the most conserved region in either of the sequences and discards the remaining sequence fragments. In cases of medium to low sequence similarity, local alignment is generally the most appropriate approach with which to start the analysis. Techniques have also been developed to align remaining sequence fragments iteratively using the local alignment technique (*e.g.*, Waterman and Eggert, 1987).

2.2.3 Type of alignment

A number of different alignment problems have been identified in the literature. For example, the BAliBASE MSA benchmark database (Thompson et al., 1999) groups these in five basic categories that contain sequence sets comprising the following features:

1. *Equidistant sequences.* Pairwise evolutionary distances between the sequences are approximately the same.

2. *Orphan sequences.* One or more family members of the sequence set are evolutionarily distant from all the others (which can be considered equidistant).

3. *Subfamilies.* Sequences are distributed over two or more divergent subfamilies.

4. *Extensions.* Alignments contain large N- and/or C-terminal gaps.

5. *Insertions.* Alignments have large internal gap insertions.

The preceding classification of alignment problems opens up the possibility of developing different alignment techniques that are optimal for each individual type of problem. Other cases that are challenging for alignment engines include repeats, where different repeat types and copy numbers often lead to incorrect alignment (see

Name	Web site
PRALINE	www.ibi.vu.nl/programs/pralinewww
MUSCLE	www.ebi.ac.uk/muscle
T-Coffee and 3D-Coffee	http://igs-server.cnrs-mrs.fr/Tcoffee/tcoffee_cgi/index.cgi
MAFFT	http://align.bmr.kyushu-u.ac.jp/mafft/online/server/
ProbCons	http://probcons.stanford.edu/
SPEM and SPEM-3D	http://sparks.informatics.iupui.edu/Softwares-Services_files/spem_3d.htm

Table 2.1: Web sites of multiple alignment programs mentioned in this chapter

Note 3 in Section 2.4), and transmembrane segments, where different hydrophobicity patterns confuse the alignment (see Note 4 in Section 2.4). However, one would then need *a priori* knowledge about the alignment problem at hand (see Note 5 in Section 2.4), which can be difficult to obtain. A suggestion for investigators is to make a first (quick) multiple alignment using general parameter settings. Often, after this first round, it becomes clear in which problem category the chosen sequence set falls, so that for further alignment parameters can be set accordingly. Remember that alignments always can be manually adjusted by using one of the available alignment editors (see Note 6 in Section 2.4).

2.3 Methods

This section highlights a selection of the most accurate MSA methods to date (Table 2.1). Each of these follows one or both of two main approaches to address the greediness of the progressive MSA protocol (see the preceding): the first is trying to avoid early match errors by using increased information for aligning pairwise sequences; the second is reconsidering alignment results and improving upon these using iterative strategies.

2.3.1 PRALINE

PRALINE is an online MSA toolkit for protein sequences. It includes a web server offering a wide range of options to optimize the alignment of input sequences, such as global or local pre-processing, predicted secondary structure information, and iteration strategies (Figure 2.2).

1. *Pre-profile processing options.* Pre-profile processing is an optimization technique used to minimize the incorporation of erroneous information during progressive alignment. The difference between this strategy and the standard global

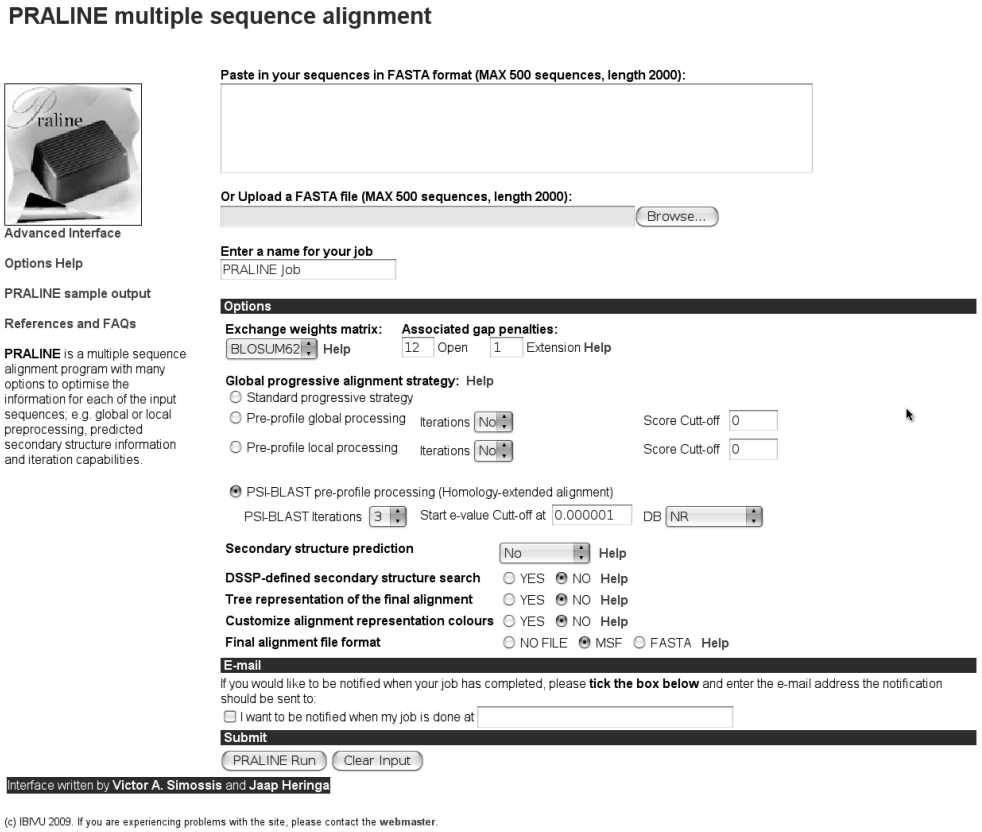


Figure 2.2: The PRALINE standard web interface. Protein sequences can be pasted in the upper box in FASTA format or directly uploaded from a file. In addition to using default settings, various alignment strategies can be selected (see Section 2.3.1) as well as the desired number of iterations or preprocessing cut-off scores.

strategy is that the sequences to be aligned are represented by pre-profiles instead of single sequences. Three different options are available: *(1)* global pre-processing (Heringa, 1999, 2002), *(2)* local pre-processing (Heringa, 2002), and *(3)* PSI-Praline (Simossis et al., 2005). The first two options attempt to maximize the information from each sequence. For each sequence, a pre-profile is built containing information from other sequences in the query set. Under global pre-processing, other sequences can be selected according to a preset minimal pairwise alignment score with the main sequence within each pre-profile. Under local pre-processing, segments of other sequences in the query set are selected based on local alignment scores. The PSI-Praline pre-profile processing strategy employs the PSI-BLAST homology search engine (Altschul et al., 1997) to enrich the information of each of the pre-profiles. Based on a user-specified E-value, PSI-BLAST selects sequence fragments from a large non-redundant sequence database, building more consistent and useful pre-profiles for the alignment. The alignment quality of the PSI-Praline strategy is among the highest in the field (Simossis et al., 2005), but the technique is relatively slow as a PSI-BLAST run needs to be conducted for every sequence in the input set.

2. *DSSP or predicted secondary structure information.* PRALINE currently allows the incorporation of DSSP-defined secondary structure information (Kabsch and Sander, 1983) to guide the alignment. If no DSSP is available, a choice of seven secondary structure prediction methods is provided to determine the putative secondary structure of those sequences that do not have a PDB structure. In addition, two different consensus strategies are also included, both relying on the prediction methods PSIPRED (Jones, 1999), PROFsec (Rost and Sander, 1993), and YASPIN (Lin et al., 2005).

3. *Iteration.* For the above global and local pre-processing strategies, iterative optimization is possible. Iteration is based on the consistency of a preceding multiple alignment, in which consistency is defined as the agreement between matched amino acids in the multiple alignment and those in corresponding pairwise alignments. These consistency scores are then fed as weights to a next round of dynamic programming. During iteration, therefore, consistent multiple alignment positions tend to be maintained, whereas inconsistent segments are more likely to become re-aligned. Iterations are terminated upon reaching convergence or limit cycle (*i.e.*, a number of cyclically recurring multiple alignments), whereas the user can also specify a maximum number of iterations.

2.3.2 MUSCLE

MUSCLE (Edgar, 2004b,c) is multiple alignment software for both nucleotide and protein sequences. It includes an online server, but the user can also choose to download the program and run it locally. The web server performs calculations using pre-defined default parameters, albeit the program provides a large number of options. MUSCLE is a very fast algorithm, which should be particularly considered

when aligning large datasets. Basically, the progressive alignment protocol is sped up due to a clever pairwise sequence comparison that avoids the slow DP technique for the construction of the so-called guide tree. Because of the computational efficiency gained, MUSCLE by default employs iterative refinement procedures that have been shown to produce high quality multiple alignments.

1. *Iteration.* The full iteration procedure used by MUSCLE consists of three steps, although only the last can be considered truly iterative.

    a. In the first step sequences are clustered according to the number of *k-mers* (contiguous segment of length k) that they share using a compressed amino acid alphabet (Edgar, 2004a). From this the guide tree is calculated using UPGMA, after which the sequences are progressively aligned following the tree order.

    b. During the next step the obtained MSA is used to construct a new tree by applying the Kimura distance correction. This step is executed at least twice and can be repeated a number of times until a new tree does not achieve any improvements anymore. As a measure to estimate improvement, the number of internal nodes for which the branching order has changed is taken. If this number remains constant or increases, the iteration procedure terminates and a last progressive alignment is built for this step.

    c. Finally, the third step involves refinement of the alignment using the now fixed tree-topology. Edges from the tree are deleted in order of decreasing distance from the root. For each subdivision of the tree, the two corresponding profiles are aligned (tree-dependent refinement step). If a resulting alignment has a higher score than the previously retained alignment, the new alignment is taken. Iteration terminates if after traversing all tree edges no new alignment is produced or the user-defined number of iterations has been reached.

2. *Large datasets.* As outlined, one of the most important advantages of MUSCLE is that it is very fast and therefore allows handling large datasets in reasonable time. A good compromise between time and accuracy can be made by the user who can decide for all stages and actions whether to include them or not. As an additional option, the user can also define a time range in which the program will select the best solution so far. Another possibility to speed up the program during pairwise k-mer alignment is provided by allowing the user to switch off extending the k-words by dynamic programming (see the preceding). A final option, called 'anchor optimization', is designed to reduce computations during tree-dependent refinement by dividing a given alignment in vertical blocks and aligning the associated profiles separately.

2.3.3 T-Coffee

The T-Coffee program (Notredame et al., 2000) can also handle both DNA and protein sequences. It includes a web server (following the default settings) as well as an option to download the program. The algorithm derives its sensitivity from combining both local and global alignment techniques. Additionally, transitivity is exploited using triplet alignment information including each possible third sequence. A pairwise alignment is created using a protocol named matrix extension that includes the following steps:

1. *Combining local and global alignment.* For each pairwise alignment, the match scores obtained from local and global alignments are summed, where for every matched residue pair the identity score of the associated (global or local) alignment is taken. For each sequence pair, the 10 highest scoring local alignments are compiled using Lalign (Huang and Miller, 1991) and a global alignment is calculated using ClustalW (Thompson et al., 1994).

2. *Transitivity.* For each third sequence C relative to a considered sequence pair A and B, the alignments A-C and C-B together constitute an alignment A-B. For each matched residue x in A and y in B, the minimum of the score of the match between residue x in A with residue z in C (alignment A-C) and that of residue z in C with y in B (alignment C-B) is taken; identity scores of associated alignments are taken as in the preceding step and all scores from the direct alignment as well as through all third sequences are summed.

3. For each sequence pair, dynamic programming is performed over the thus extended matrices. Owing to the fact that the signal captured in the extended scores is generally more consistent than noise, the scores are generally salient such that gap penalties can be set to zero.

   From the extended alignment scores a guide tree is calculated using the Neighbor-Joining technique, and sequences are progressively aligned following the dynamic programming protocol. The combined use of local alignment, global alignment, and transitivity effectively alleviates error propagation during progressive alignment. However, the program is constrained by computational demands when aligning larger sets. As a consequence, the T-Coffee web server constrains the allowed number of input sequences to 50. T-Coffee permits the following further features:

4. *Integrating tertiary structures with 3D-Coffee.* A variant of the described protocol, 3D-Coffee (O'Sullivan et al., 2004) allows the inclusion of tertiary structures associated with one or more of the input sequences for guiding the alignment based upon the principle that 'Structure is more conserved than sequence.' If a partial sequence of a structure is given, the program will only take the corresponding structural fragment into account. The 3D-Coffee web server incorporates two default pairwise structural alignment methods: SAP (Taylor and

Orengo, 1989) and FUGUE (Shi et al., 2001). The first method is a structure superposition package, which is useful if more than one structure is included. The latter is a threading technique that can improve the multiple alignment process when local structural fragments are available. The advanced interface of the program allows the user to select alternative structural alignment methods.

5. *Accelerating the analyses.* Speed limitations of the T-Coffee program can be partially reduced by running a less demanding version. As an alternative, sequences can be divided into subgroups and aligned separately. To assist in this scenario, the program offers an option to compile a final alignment of these previously aligned subgroups.

6. *Consensus MSA.* A recent extension is the method M-Coffee (Wallace et al., 2006), which uses the T-Coffee protocol to combine the outputs of other MSA methods into a single consensus MSA.

2.3.4 MAFFT

The multiple sequence alignment package MAFFT (Katoh et al., 2002, 2005) is suited for DNA and protein sequences. MAFFT includes a script and a web server that both incorporate several alignment strategies. An alternative solution is proposed for the construction of the guide tree, which usually requires most computing time in a progressive alignment routine. Instead of performing all-against-all pairwise alignments, Fast Fourier Transformation (FFT) is used to rapidly detect homologous segments. The amino acids are represented by volume and polarity values, yielding high FFT peaks in a pairwise comparison whenever homologous segments are identified. The segments thus identified are then merged into a final alignment by dynamic programming. Additional iterative refinement processes, in which the scoring system is quickly optimized at each cycle, yield high accuracy of the alignments.

1. *Fast alignment strategies.* Two options are provided for large sequence sets: FFT-NS-1 and FFT-NS-2, both of which follow a strictly progressive protocol. FFT-NS-1 generates a quick and dirty guide tree and compiles a corresponding MSA. If FFT-NS-2 is invoked, it takes the alignment obtained by FFT-NS-1 but now calculates a more reliable guide tree, which is used to compile another MSA.

2. *Iterative strategies.* The user can choose from several iterative approaches. The FFT-NS-i method attempts to further refine the alignment obtained by FFT-NS-2 by re-aligning subgroups until the maximum weighted sum of pairs (WSP) score (Gotoh, 1995) is reached. Two more recently included iterative refinement options (MAFFT version 5.66) incorporate local pairwise alignment information into the objective function (sum of the WSP scores). These are L-INS-i and E-INS-i, which use standard affine and generalized affine gap costs (Altschul, 1998; Zachariah et al., 2005) for scoring the pairwise comparisons, respectively.

3. *Alignment extension.* Another tool included in the MAFFT alignment package is mafftE. This option enhances the original dimension of the input set by including other homologous sequences, retrieved from the SwissProt database with BLAST (Altschul et al., 1990). Preferences for the exact number of additional sequences and the e-value can be specified by the user.

2.3.5 ProbCons

ProbCons (Do et al., 2005) is a recently developed progressive alignment algorithm for protein sequences. The software can be downloaded but sequences can also be submitted to the ProbCons web server. The method follows the T-Coffee approach in spirit, but implements some of the steps differently. For example, the method uses an alternative scoring system for pairs of aligned sequences. The method starts by using a pair-HMM and expectation maximization (EM) to calculate a posterior probability for each possible residue match within a pairwise comparison. Next, for each pairwise sequence comparison, the alignment that maximizes the 'expected accuracy' is determined ((Holmes and Durbin, 1998). In a similar way to the T-Coffee algorithm, information of pairwise alignments is then extended by considering consistency with all possible third 'intermediate' sequences. For each pairwise sequence comparison, this leads to a so-called 'probabilistic consistency' that is calculated for each aligned residue pair using matrix multiplication. These changed probabilities for matching residue pairs are then used to determine the final pairwise alignment by dynamic programming. Upon construction of a guide tree, a progressive protocol is followed to build the final alignment.

ProbCons allows a few variations of the protocol that the user can decide to adopt:

1. *Consistency replication.* The program allows the user to repeat the probabilistic consistency transformation step, by recalculating all posterior probability matrices. The default setting includes two replications, which can be increased to a maximum of 5.

2. *Iterative refinement.* The program also includes an additional iterative refinement procedure for further improving alignment accuracy. This is based on repeated random subdivision of the alignment in two blocks of sequences and realignment of the associated profiles. The default number of replications is set to 100, but can be changed from 0 to 1000 iterations (for the web server one can select 0, 100, or 500).

3. *Pre-training.* Parameters for the pair-HMM are estimated using unsupervised expectation maximization (EM). Emission probabilities, which reflect substitution scores from the BLOSUM-62 matrix (Henikoff and Henikoff, 1992), are fixed, whereas gap penalties (transition probabilities) can be trained on the whole set of sequences. The user can specify the number of rounds of EM to be applied on the set of sequences being aligned. The default number of iterations should be followed, unless there is a clear need to optimize gap penalties when considering a particular dataset.

2.3.6 SPEM

The SPEM-protocol (Zhou and Zhou, 2005), designed for protein MSA, is a recent arrival in the field. Both a SPEM server and downloadable software are available. Two online SPEM protocols are available: SPEM (normal) and SPEM-3D. Each follows a standard routine so that the user cannot change many options. The 3D-variant SPEM-3D, which allows the inclusion of information from tertiary structure, can only be used through the Web. The SPEM approach focuses on the construction of proper pairwise alignments, which constitute the input for the progressive algorithm. To optimize pairwise alignment, the method follows the PRALINE approach (see the preceding) in that it combines information coming from sequence pre-profiles (constructed *a priori* with homology searches performed by PSI-BLAST) (Altschul et al., 1997), and knowledge about predicted and known secondary structures. However, the latter knowledge is exploited in the dynamic programming algorithm by applying secondary structure dependent gap penalty values, whereas PRALINE in addition uses secondary structure-specific residue exchange matrices. The pairwise alignments are further refined by a consistency-based scoring function that is modeled after the T-Coffee scenario (see the preceding) based on integrating information coming from comparisons with all possible third sequences.

Next, a guide tree is calculated based on sequence identities and followed to determine the progressive multiple alignment path, leading to a final MSA based on the refined pairwise alignments. The web servers for SPEM and SPEM-3D can handle up to 200 sequences, whereas for the 3D version maximally 100 additional structures can be included.

2.4 Notes

1. *Distant sequences*: able to make very accurate MSAs, alignment incompatibilities can arise under divergent evolution. In practice, it has been shown that the accuracy of all alignment methods decreases dramatically whenever a considered sequence shares <30% sequence identity (Rost, 1999). Given this limitation, it is advisable to compile a number of MSAs using different amino acid substitution matrices. Among these, the PAM (Dayhoff et al., 1978) and BLOSUM (Henikoff and Henikoff, 1992) series of substitution matrices are the most widely used (especially BLOSUM62). It is helpful to know that higher PAM numbers and low BLOSUM numbers (*e.g.*, PAM250 or BLOSUM45) correspond to exchange matrices that have been designed for the alignment of increasingly divergent sequences, respectively, whereas matrices with lower PAM and higher BLOSUM numbers are suitable for more closely related sequence sets. Furthermore, it is crucial to attempt different gap penalty values, as these can greatly affect the alignment quality. Gap penalties are an essential part of protein sequence alignment when using dynamic programming. The higher the gap penalties, the stricter the insertion of gaps into the alignment and consequently the fewer gaps inserted. Gap regions in an MSA often correspond to loop re-

gions in the associated tertiary structures, which are preferentially altered by divergent evolution. Therefore, it can be useful to lower the gap penalty values for more divergent sequence sets, although care should be taken not to deviate too much from the recommended settings. Excessive gap penalty values will enforce a gap-less alignment, whereas low gap penalties will lead to alignments with very many gaps, allowing (near) identical amino acids to be matched. In both cases the resulting alignment will be biologically inaccurate. The way in which gap penalties affect the alignment also depends on the residue exchange matrix used. Although recommended combinations of exchange matrices and gap penalties have been described in the literature and most methods include default matrices and gap penalty settings, there is no formal theory yet as to how gap penalties should be chosen given a particular residue exchange matrix. Therefore, gap penalties are set empirically: for example, penalties of 11 and 1 are recommended for BLOSUM62, whereas the suggested values for PAM250 are 10 and 1.

2. *Multi-domain proteins (Dialign, T-Coffee)*: Multi-domain proteins can be a particular challenge for multiple alignment methods. Whenever there has been an evolutionary change in the domain order of the query protein sequences, or if some domains have been inserted or deleted across the sequences, this leads to serious problems for global alignment engines. Global methods are not able to deal with permuted domain orders and normally exploit gap penalty regimes that make it difficult to insert long gaps corresponding to the length of one or more protein domains. For the alignment of multi-domain protein sequences, it is advisable to resort to a local multiple alignment method. Alternatively, the TCoffee (Notredame et al., 2000) and Dialign (Morgenstern et al., 1996; Morgenstern, 2004) methods might provide a meaningful alignment of multi-domain proteins, as they are (partly) based on the local alignment technique.

3. *Repeats*: The occurrence of repeats in many sequences can seriously compromise the accuracy of MSA methods, mostly because the techniques are not able to deal with different repeat copy numbers. Recently, an MSA strategy has become available that keeps track of various repeat types (Sammeth and Heringa, 2006). The method requires the specification of the individual repeats, which can be obtained by running one of the available repeat detection algorithms, after which a repeat-aware MSA is produced. Although the alignment result can be markedly improved by this method, it is sensitive to the accuracy of the repeats information provided.

4. *TM regions*: A special class of proteins is comprised of membrane-associated proteins. The regions within such proteins that are inserted in the cell membrane display a profoundly changed hydrophobicity pattern as compared with soluble proteins. Because the scoring schemes (*e.g.*, PAM (Dayhoff et al., 1978) or BLOSUM (Henikoff and Henikoff, 1992)) normally used in MSA techniques

are derived using sequences of soluble proteins, the alignment methods are in principle not suitable to align membrane bound protein regions. This means that great care should be taken when using general MSA methods. Fortunately, transmembrane (TM) regions can be reliably recognized using state-of-the-art prediction techniques such as TMHMM (Krogh et al., 2001) or Phobius (Käll et al., 2004). Therefore, it can be advisable to mark the putative TM regions across the query sequences, and if their mutual correspondence would be clear, to align the blocks of intervening sequence fragments separately.

5. *Preconceived knowledge*: In many cases, there is already some preconceived knowledge about the final alignment. For instance, consider a protein family containing a disulfide bridge between two specific cysteine residues. Given the structural importance of a disulfide bond, constituent Cys residues are generally conserved, so that it is important that the final MSA matches such Cys residues correctly. However, depending on conservation patterns and overall evolutionary distances of the sequences, it can well happen that the alignment engine needs special guidance for matching the Cys residues correctly. Currently none of the approaches has a built-in tool to mark particular positions and assign specific parameters for their consistency, although the library structure of the T-Coffee method allows the specification of weights for matching individual amino acids across the input sequences. However, exploiting this possibility can be rather cumbersome. The following suggestions are therefore offered for (partially) resolving this type of problem:

    a. *Chopping alignments.* Instead of aligning whole sequences, one can decide to chop the alignment in different parts. For example, this could be done if the sequences have some known domains for which the sequence boundaries are known. An added advantage in such cases is that no undesirable overlaps will occur between these pre-marked regions if aligned separately. Finally, the whole alignment can be built by concatenating the aligned blocks. It should be stressed that each of the separate alignment operations is likely to follow a different evolutionary scenario, as for example the guide tree or the additionally homologous background sequences in the PSI-PRALINE protocol can well be different in each case. It is entirely possible, however, that these different scenarios reflect true evolutionary differences, such as for instance unequal rates of evolution of the constituent domains. In the first step sequences are clustered according to the number of *k-mers*

    b. *Altering amino acid exchange weights.* Multiple alignment programs make use of amino acid substitution matrices in order to score alignments. Therefore, it is possible to change individual amino acid exchange values in a substitution matrix. Referring to the disulfide example mentioned in the preceding, one could decide to up-weight the substitution score for a cysteine self-conservation. As a result, the alignment will obtain a higher

score when cysteines are matched, and as a consequence the method will attempt to create an alignment where this is the case. However, some protein families have a number of known pairs of Cys residues that form disulfide bonds, where mixing up of the Cys residues involved in different disulfide bridges might happen in that Cys residues involved in different disulfide bonds become aligned at a given single position. To avoid such incorrect matches in the alignment, some programs (*e.g.*, PRALINE) allow the addition of a few extra amino acid designators in the amino acid exchange matrix that can be used to identify Cys residue pairs in a given bond (*e.g.*, J, O, or U). The exchange scores involving these 'alternative' Cys residues should be identical to those for the original Cys, except for the cross-scores between the alternative letters for Cys that should be given low (or extreme negative) values to avoid cross alignment. It must be stressed that such alterations are heuristics that can violate the evolutionary model underlying a given residue exchange matrix.

6. *Alignment editors*: A number of multiple alignment editors are available for editing automatically generated alignments, which often can be improved manually. Posterior manual adjustments can be helpful, especially if structural or functional knowledge of the sequence set is at hand. The following editing tools are available:

   a. *Jalview* (`www.jalview.org`) (Clamp et al., 2004) is a protein multiple sequence alignment editor written in Java. In addition to a number of editing options, it also provides a wide scale of sequence analysis tools, such as sequence conservation, UPGMA, and NJ (Saitou and Nei, 1987) tree calculation, and removal of redundant sequences. Color schemes can also be customized according to amino acid physiochemical properties, similarity to consensus sequence, hydrophobicity, or secondary structure.

   b. *SeaView* (`http://pbil.univ-lyon1.fr/software/seaview.htm`) (Galtier et al., 1996) is a graphical editor suited for Mac, Windows, Unix, and Linux. The program includes a dot-plot routine for pairwise sequence comparison (Li and Graur, 1991) or the ClustalW (Thompson et al., 1994) multiple alignment program to locally improve the alignment and can also perform phylogenetic analyses. Again, color schemes can be customized.

   c. *STRAP* (`www.charite.de/bioinf/strap/`) (Gille and Frömmel, 2001) is an interactively extendable and scriptable editor program, able to manipulate large protein alignments. The software is written in Java and is compatible with all operating systems. Among the many extra features provided are: enhanced alignment of low-similarity sequences by integrating 3D-structure information, determination of regular expression motifs, and transmembrane and secondary structure predictions.

d. *CINEMA* (`www.bioinf.manchester.ac.uk/dbbrowser/CINEMA2.1/`) (Parry-Smith et al., 1998) is a Java interactive tool for editing either nucleotide or amino acid sequences. The flexible editor permits color scheme changes and motif selection. Hydrophobicity patterns can also be viewed. Furthermore, there is an option to load prepared alignments from the PRINTS fingerprint database (Attwood et al., 1997).

CHAPTER 3

PRALINE^{TM}: a strategy for improved multiple alignment of transmembrane proteins

Published as:

Pirovano, W., Feenstra, K.A., and Heringa, J. (2008).
PRALINE^{TM}: a strategy for improved multiple alignment of transmembrane proteins.
Bioinformatics, 24(4):492–497.

Abstract

Background: Membrane-bound proteins are a special class of proteins. The regions that insert into the cell-membrane have a profoundly different hydrophobicity pattern compared with soluble proteins. Multiple alignment techniques use scoring schemes tailored for sequences of soluble proteins and are therefore in principle not optimal to align membrane-bound proteins.
Results: Transmembrane (TM) regions in protein sequences can be reliably recognized using state-of-the-art sequence prediction techniques. Furthermore, membrane-specific scoring matrices are available. We have developed a new alignment method, called PRALINE^{TM}, which integrates these two features to enhance multiple sequence alignment. We tested our algorithm on the TM alignment benchmark set by Bahr et al. (2001), and showed that the quality of TM alignments can be significantly improved compared with the quality produced by a standard multiple alignment technique. The results clearly indicate that the incorporation of these new elements into current state-of-the-art alignment methods is crucial for optimizing the alignment of TM proteins.
Availability: A webserver is available at `www.ibi.vu.nl/programs/pralinewww`.

3.1 Introduction

Over the past years, integral membrane proteins have received a great deal of attention. They carry out essential functions in many cellular and physiological processes, such as signal transduction, cellcell recognition and molecular transport. Membrane proteins are likely to constitute 20-30% of all ORFs contained in genomes (Jones, 1998; Wallin and von Heijne, 1998).

Unfortunately, the number of determined transmembrane (TM) structures in the PDB is still very low: <2% of all structures solved show a membrane topology (`www.pdb.org`; Tusnády et al., 2005). Despite of a solid exponential growth of the number of membrane protein structures (White, 2004), their determination remains a difficult task, such that they will continue to lag behind relative to the number of elucidated soluble protein structures.

Transmembrane (TM) regions show a modified hydrophobicity and conservation pattern as compared with soluble proteins. Conventional scoring matrices such as PAM (Dayhoff et al., 1978) or BLOSUM (Henikoff and Henikoff, 1992), routinely used for sequence retrieval and alignment, are therefore in principle not suitable to align membrane-bound protein regions. Jones et al. (1994b) for instance noticed that polar residues are highly conserved in these regions, whereas hydrophobic residues are more interchangeable, and developed the JTT TM substitution matrix. Ng et al. (2000) derived a new TM-specific substitution matrix called PHAT, which was shown to outperform the JTT matrix, especially on database searching (Ng et al., 2000). Meanwhile several groups focused on the development of accurate membrane topology predictors such as HMMTOP (Tusnády and Simon, 1998, 2001), TMHMM (Krogh

et al., 2001; Sonnhammer et al., 1998), Phobius (Käll et al., 2004, 2005) and MEMSAT (Jones, 2007; Jones et al., 1994a). The topic has recently been reviewed by Punta et al. (2007).

Not many techniques however have been developed to improve the alignment of TM proteins. The method STAM (Shafrir and Guy, 2004) represents an early attempt to improve alignment accuracy by combining different substitution matrices. A more recent study by Forrest et al. (2006) reported that the use of a bipartite scheme (consisting of BLOSUM62 and PHAT) does not significantly improve membrane protein sequence alignments. They suggest that the previously reported progress is more likely to depend on the separation of the TM blocks or on the settings of specific gap penalties.

In this study we have investigated the effects of incorporating TM specific information into the previously developed multiple alignment tool PRALINE (Heringa, 1999, 2002). This information is integrated in a 'soft' way, compared with for instance the STAM approach where TM segments are first chopped and then aligned separately. In our approach the choice of the matrix depends on consistent TM predictions over a column and is determined dynamically during the alignment procedure. We also explore an additional iterative strategy to further optimize the alignments.

We have tested the algorithm on the TM benchmark alignments of BAliBASE (Bahr et al., 2001). This reference set contains more than 400 reliably aligned TM sequences divided into eight families. The alignments are manually curated and at the moment they constitute by far the largest available benchmark. By applying the PHAT substitution matrix on accurately predicted TM regions combined with a proper gap penalty setting, we show that we are able to significantly improve the alignment quality.

3.2 Methods

The 'basic' and 'global profile pre-processing' ('pre-profile' or 'prepro') PRALINE progressive alignment algorithms, underlying the strategies tested in this study, are described in detail in previously published works (Heringa, 1999, 2002). In brief, the 'basic' PRALINE alignment method simply follows the classic progressive alignment protocol where sequences are aligned following the order of the guide tree. In the 'pre-profile' method for each sequence a so-called master-slave alignment is constructed, containing information about neighboring sequences, which are then used in subsequent progressive alignment. It has been shown that these sequence pre-profiles are more informative than single sequences and help to avoid mistakes during the progressive steps (Heringa, 2002). For the PRALINE^{TM} tool we present here, we first predict for each input sequence its TM topology using a state-of-the-art predictor. Second, the profile-scoring scheme simply applies TM-specific substitution scores from the PHAT matrix to reliably predicted TM positions. Finally, we incorporated an alternative iterative scheme to enhance the alignment quality. In Figure 3.1 an overview of the PRALINE^{TM} strategy is given.

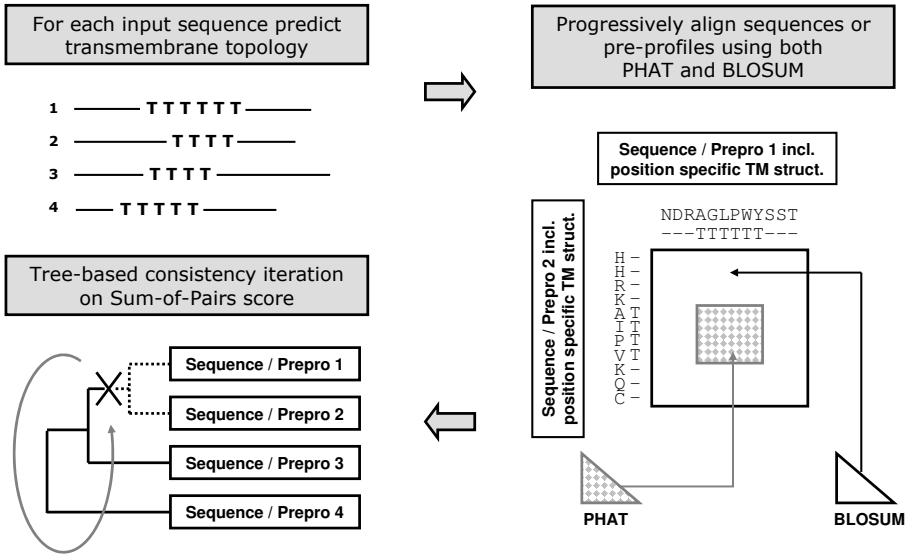


Figure 3.1: Overview of the PRALINE^{TM} strategy.

3.2.1 Scoring scheme

The current PRALINE profile-scoring scheme uses the following equation to score a pair of profile columns x and y:

$$S(a,b) = \sum_{i}^{20} \sum_{j}^{20} \alpha_i \beta_j M(i,j) \tag{3.1}$$

where α_i and β_j are the frequencies with which residues i and j appear in columns x and y, respectively, and $M(i,j)$ is the exchange weight for residues i and j provided by the selected substitution matrix M. By default profile columns are aligned using the BLOSUM62 matrix. Two profile columns will be matched using the PHAT matrix only in case each residue in the column is predicted to be member of a TM segment. This is done to guarantee that inconsistently predicted positions do not negatively influence the alignment quality. As a result, and contrary to the STAM method (Shafrir and Guy, 2004), our approach potentially allows TM segments to be aligned to non-TM segments. The BLOSUM62 and PHAT substitution matrices are normalized using their diagonal elements as described in (Abagyan and Batalov, 1997).

3.2.2 Transmembrane topology predictors

Transmembrane topologies are predicted using three different state-of-the-art methods: HMMTOP v2.1 (Tusnády and Simon, 2001), TMHMM v2.0 (Krogh et al., 2001)

and Phobius (Käll et al., 2004). All predictors are installed locally and run independently within the PRALINE^{TM} program.

3.2.3 Multiple alignment benchmark

To evaluate the performance of the two standard PRALINE alignment techniques, the PRALINE^{TM} method and other state-of-the-art multiple alignment programs, we used the BAliBASE (v2.0) reference alignment set of TM proteins (Bahr et al., 2001). The set includes eight accurately aligned TM families. The total number of sequences is 435 with an average length of 567 residues. The number of TM helices per sequence varies from 2 to 14.

The BAliBASE 'testing' program was used to evaluate alignment accuracy against the benchmark. Accuracy is measured with two alternative scores: the 'SP score' measures the number of correctly aligned residue pairs in the test alignment to the number of aligned pairs in the reference alignment, while the 'TC score' measures the number of correctly aligned columns in the test alignment to the number of columns in the reference alignment.

3.2.4 Alignment methods tested

First of all we tested the performance of the original PRALINE tools compared with the PRALINE^{TM} application where two matrices are combined. It is commonly thought that gaps within TM regions should be more penalized than gaps in soluble regions. We therefore evaluated different combinations of gap penalties, to see whether the sensitivity of the approach resides in the gap penalty settings or the specific TM matrix. In addition we compared the results obtained to other multiple alignment routines, which are designed for standard alignment purposes. These include: ClustalW v1.83 (Thompson et al., 1994), MUSCLE v3.52 (Edgar, 2004b,c), MAFFT v6 (Katoh et al., 2005) and ProbCons v1.12 (Do et al., 2005) and all of these programs were run using default parameter settings.

3.2.5 Tree-based consistency iteration

We also employed the potential benefits of an additional iterative strategy. At the heart of it lies the tree-dependent consistency iteration, which is similar to the tree-dependent strategy proposed by Hirosawa et al. (1995) and its implementation in the MUSCLE method (Edgar, 2004b,c). In this scenario each edge of the phylogenetic (guide) tree is used to divide the alignment into two sub-alignments, which are successively realigned. The new alignment is retained only if a higher Sum-of-Pairs score is achieved. In our case this score is obtained by summing the substitution values of both the BLOSUM62 and PHAT matrix (depending on the TM topology of the amino acid pair). For the tree-based consistency strategy one iterative cycle means that each edge of the tree is visited once. The maximum number of iterations is set to 20.

Method	SP score	TC score
PRALINE basic	0.646	0.231
PRALINE^{TM} basic – HMMTOP	0.679	0.264
PRALINE^{TM} basic – TMHMM	0.725	0.254
PRALINE^{TM} basic – Phobius	0.737	0.268

Table 3.1: Performance of the PRALINE and PRALINE^{TM} basic strategies on reference set 7 of BAliBASE (at gap-open and gap-extension penalties of 15.0 and 1.0 for both the soluble and the transmembrane regions)

3.3 Results and discussion

3.3.1 Performance of the PRALINE^{TM} methods compared with the standard PRALINE methods

First of all we sought to understand whether a general improvement of alignment quality could be observed when including the TM-specific information. We therefore extensively tested both 'PRALINE basic' and 'PRALINE prepro' using the three selected TM topology predictors (see Section 3.2) and gapopen penalties ranging from 12 to 18 (in steps of 1) for both the soluble and the TM regions. An additional parameter, the pre-profile cut-off, was varied from 8.0 to 15.0 (in steps of 0.5) following the global pre-processing conditions defined in Heringa (2002). This parameter indicates to what extent other neighboring sequences are included into the sequence pre-profiles. The results of the 'PRALINE basic' strategies are summarized in Table 3.1, the results of the 'PRALINE prepro' strategies in Figure 3.2 and Figure 3.3.

The most striking observation to be made from both Table 3.1 and Figure 3.2a and b is the positive effect on the alignment quality of the PHAT matrix applied on reliably predicted TM regions. Here the results are shown at an arbitrary gap-open penalty of 15.0 and gap-extension penalty of 1.0 for both the soluble and the TM regions; the outcomes are consistent over all combinations of gap-open penalties.

A notable increase can be observed for all three TM predictors, albeit Phobius gives the best performance overall. Phobius has shown to be one of the most accurate TM topology predictors, especially on sequences that also contain a signal peptide (Jones, 2007; Käll et al., 2004).

Concerning the pre-profile cut-off it can be noticed from Figure 3.2a and b that the optimal parameter settings lie between 11.0 and 12.0. In this range the highest SP and TC scores are reached and also maximum improvement relative to the standard pre-profile technique is attained. Consistency of these scores was estimated by 8-fold cross-validation, each time leaving out one BAliBASE alignment and retaining the other seven (data not shown). Standard deviation over the cross-validated SP scores was below 4%, and minimal SD of around 1.5% were reached at highest SP score. For

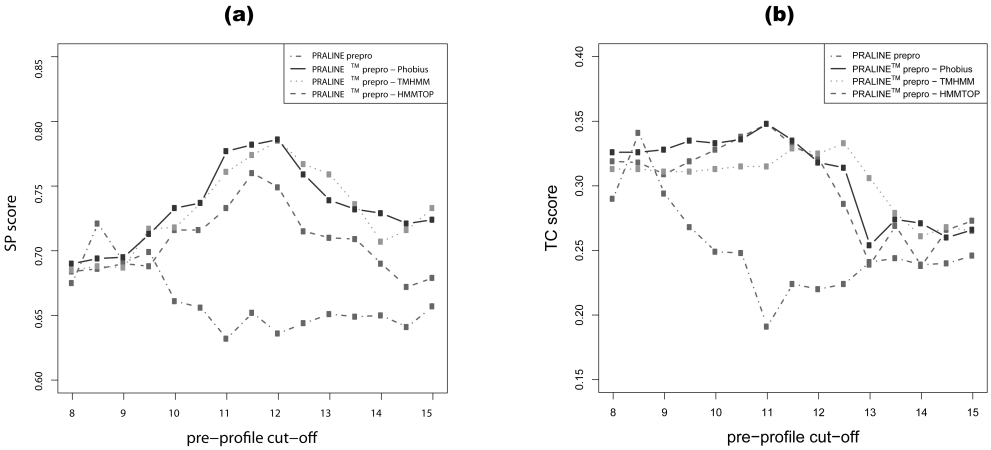


Figure 3.2: Performance of the PRALINE and PRALINE^{TM} prepro strategies on reference set 7 of BAliBASE (gap settings as for Table 3.1. In **(a)** the average SP score is plotted, in **(b)** the average TC score.

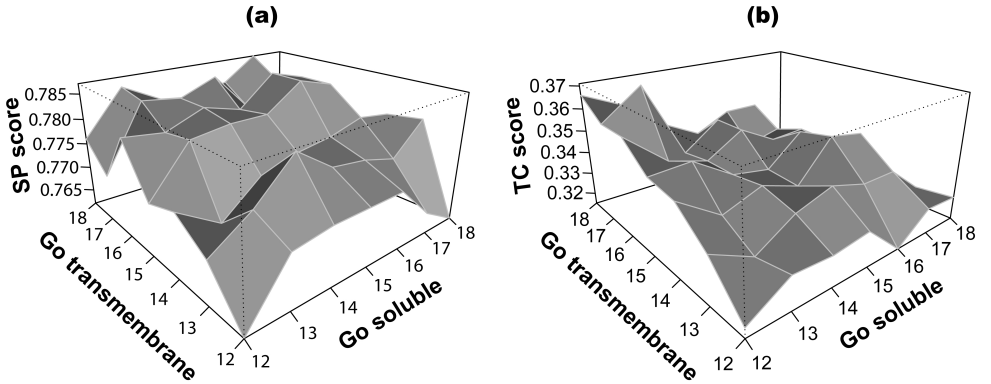


Figure 3.3: Effects of different gap-open ('Go') penalty combinations on the performance of the 'PRALINE^{TM} prepro - Phobius method' at a mid-range pre-profile cut-off of 11.5 Figure 3.2. In **(a)** the results of the SP score measure are shown, in **(b)** the results of the TC score measure.

the TC scores, SD were between 3 and 4.5%; here the differences are small enough that the optimal choice also lies at the highest TC scores.

Finally, Figure 3.3 shows the performance of the 'PRALINE^{TM} prepro – Phobius' method at different gap-open penalties: for the soluble and TM regions different penalty combinations were tested. An optimal gap-open penalty for the soluble regions is hard to define, though the optimum for TM regions lies between 15.0 and 18.0. We also varied both the soluble and TM gap-extension penalties from 1.0 to 1.5, but no significant differences were observed.

3.3.2 Iterative strategies further improve the PRALINE^{TM} method

Next, we studied the effects of the additional tree-based consistency iterative strategy (described in Section 3.2.5 on the PRALINE^{TM} method. Following our above results we now varied the pre-profile cut-off between 11.0 and 12.0, the gap-open penalties from 12.0 to 18.0 for soluble regions (in steps of 1.0) and from 15.0 to 18.0 for TM regions (in steps of 0.5). Over the whole range of settings the iterative procedure improved the PRALINE^{TM} outcomes by $\sim$1.2 percentage points using SP and TC scoring. Specifically, the optimal parameter setting found was at a pre-profile cut-off of 11.0 combined with a gap-open penalty combination of 15.0 for the soluble regions and 16.5 for the TM regions. The above mentioned parameter settings define the final PRALINE^{TM} method.

3.3.3 Contributions of substitution matrices and gap penalty settings on the alignment quality

We also investigated independent contributions to the alignment quality coming from the PHAT matrix and specific gap penalties. For this purpose, we tested PRALINE^{TM}, defined in the previous section, at different gap-open penalties for the TM regions. In the first run the TM segments were aligned using the PHAT matrix. In the other two runs we used either the standard BLOSUM62 matrix or the PHAT matrix for the entire sequence. The results in Figure 3.4 clearly show that only the combination of BLOSUM62/PHAT matrices yield optimal results. The runs in which only one matrix is applied to the whole sequence, even when optimized gap penalties are used, produce much less reliable alignments. On the other hand, we noticed that applying a slightly higher gap-open penalty to the TM regions relative to that for soluble regions can have some additional benefits. These influences however are much less pronounced, implying that the results obtained are not very sensitive to the TM gap-open settings.

We further tested the BLOSUM55 matrix, since it has an entropy comparable to that of the PHAT matrix ($H = 0.5637$) and $H = 0.5605$, respectively). We optimized the alignment parameters for the BLOSUM55/PHAT combination as in Section 3.3.2 and observed a decreased performance of 5 percentage points on average. A likely explanation of this effect can be that the evolutionary scenarios underlying TM and soluble regions are different.

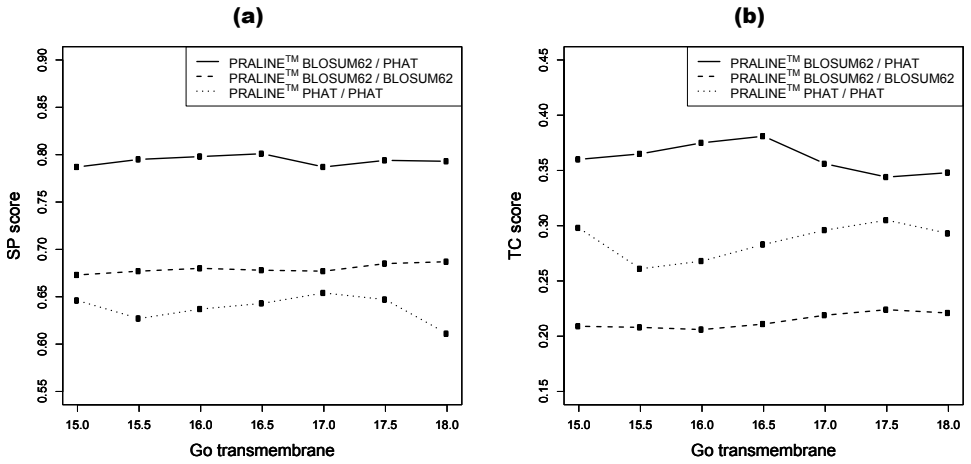


Figure 3.4: Contributions of the PHAT matrix and the gap penalty settings to the alignment quality. The final PRALINE^{TM} method is evaluated at different gap-open penalty values for the transmembrane regions. In the first analysis (BLOSUM62/PHAT) we apply the PHAT matrix to the transmembrane regions. In the two control analyses we apply either BLOSUM62 (BLOSUM62/BLOSUM62) or PHAT (PHAT/PHAT) to the entire sequence. In **(a)** the average SP score is plotted, in **(b)** the average TC score.

3.3.4 Comparison with other alignment methods

Finally we compared our algorithm with widely-used multiple alignment methods, which are designed for aligning soluble proteins. Results are shown in Table 3.2. The standard PRALINE (*i.e.* 'prepro' without TM information) method with optimized parameter settings over this dataset is included for reference (at a gap-open penalty of 15.0 the optimal pre-profile cut-off is 8.5, see Figure 3.2). Notably, all methods reach SP scores that are twice as high as corresponding TC scores. The latter score is a much stricter measure, but arguably also more meaningful since evolutionary analysis is usually performed on whole alignment columns. We see that PRALINE^{TM} achieves the highest SP score for two datasets and the highest TC score for four datasets. Concerning the averages over all eight datasets, ProbCons slightly outperforms MAFFT (-0.6 percentage points) and PRALINE^{TM} (-0.5 percentage points) on the SP score. On the more critical TC score PRALINE^{TM} clearly scores best ($+1.5$ and $+5.1$ percentage points compared with ProbCons and MAFFT, respectively). ClustalW and MUSCLE score considerably lower on almost all datasets. The standard PRALINE method achieves a SP score comparable to ClustalW, but can be placed between MAFFT and ProbCons with respect to the TC score. The inclusion of TM information in PRALINE^{TM} yields $+8.0$ percentage points for SP scoring and

Set	ClustalW	MUSCLE	MAFFT	ProbCons	PRALINE	PRALINE^{TM}
SP score						
7tm	0.847	0.836	0.835	**0.882**	0.816	0.860
acr	0.906	**0.946**	0.937	0.935	0.930	0.936
dtd	0.786	0.855	0.844	**0.877**	0.824	0.863
ion	0.354	0.520	0.509	0.527	0.346	**0.543**
msl	0.864	0.870	0.845	0.849	0.813	**0.874**
Nat	0.630	0.738	**0.766**	0.745	0.720	0.713
photo	0.887	0.902	**0.934**	0.913	0.915	0.933
ptga	0.461	0.551	**0.729**	0.716	0.404	0.683
AVG	0.717	0.777	0.800	**0.806**	0.721	0.801
TC score						
7tm	0.410	0.340	0.320	0.410	0.310	**0.430**
acr	0.580	0.670	0.620	0.670	**0.690**	0.620
dtd	0.250	0.310	0.210	0.340	0.360	**0.390**
ion	0.000	0.000	0.030	**0.090**	0.000	0.000
msl	0.610	0.630	0.610	0.600	0.580	**0.660**
Nat	0.020	0.130	0.120	0.180	**0.220**	0.140
photo	0.490	0.460	0.550	0.490	0.570	**0.730**
ptga	0.010	0.060	**0.180**	0.150	0.000	0.080
AVG	0.296	0.325	0.330	0.366	0.341	**0.381**

Table 3.2: Comparison between the PRALINE^{TM} method and four widely-used multiple alignment methods. Also the best scoring PRALINE ('prepro') method without TM information is included. Individual and average SP and TC scores are given; for each set the best scoring method is highlighted in bold.

+4.0 percentage points for TC scoring compared with standard PRALINE.

It should be stressed that the PRALINE and PRALINE^{TM} methods were optimized on the TM dataset, whereas the other methods were run at default settings. Concerning this, both MAFFT and ProbCons are relatively robust on TM sequences. Nonetheless, the results show clearly that our TM-based strategy can significantly improve the quality of TM protein sequence alignments, and should be considered a promising avenue for other applications as well.

3.4 Conclusions

We present a new strategy designed to accurately align protein families adopting a TM topology. We conclude that the alignment quality can be improved significantly using a TM-specific substitution matrix and proper gap penalty settings. In our view the improvement is mainly attributed by the fact that the bipartite scheme, using BLOSUM62 and PHAT, is applied in a flexible manner to undivided sequences during each step of the alignment procedure. To the best of our knowledge, the magnitude of the success accomplished has not been reported elsewhere to date. Other attempts where TM and soluble regions were aligned independently did not succeed in making

significantly better alignments (Forrest et al., 2006). In fact, in those approaches the definition of the TM segment is of crucial importance as TM segments cannot be aligned with non-TM segments, such that incorrectly delineated TM regions are likely to lead to misaligned TM and soluble segments. Even bigger problems arise when the number of TM segments varies within families. PRALINE^{TM} aligns undivided sequences instead and applies substitution scores from the PHAT matrix only where predictions are 100% consistent. The flexibility of the algorithm allows TM segments to be aligned with non-TM segments if other signals prevail over the TM signals.

It should be stressed that the choice of the prediction method can play an important role. Although in this article we did not explicitly test the quality of the different methods, their specific algorithms certainly affect the alignment outcomes. In general, all three methods tested enhanced the alignment accuracy, while Phobius emerged as the most valuable tool for TM alignment. Phobius is considered one of the most accurate TM topology predictors and its main advantage resides in the ability to discriminate between TM segments and signal peptides (Jones, 2007; Käll et al., 2004). The fact that most BAliBASE reference alignments contain predicted signal peptides explains at least partly the leading role of Phobius when used in our alignment strategy.

It is reassuring that the alignment quality we obtain with PRALINE^{TM} is correlated with the prediction quality of the TM prediction methods reported in the literature for TM sequences containing signal peptides (Jones, 2007; Käll et al., 2004). Further improvements could come from incorporating prediction confidence levels, or combining the TM topology predictions in a single consensus prediction.

Concerning gap penalties we noticed that strict gap penalty settings for TM regions improve the overall performance. However, these effects should not be overestimated: we found that the optimal gap-open penalty applied to the TM segments was only about 10% higher than the standard gap-open penalty applied to soluble regions.

None of the methods included here was able to align more than 40% of the reference alignment columns on average, so that further optimization remains a challenging task. Nonetheless this research has shed some new light on the alignment of TM protein families and shows that TM-awareness is an important concept for optimizing multiple sequence alignment quality.

3.5 Acknowledgements

The authors wish to thank Thomas W. Binsl for invaluable discussions and help with the implementation. We are also grateful to Gabor E. Tusnády, Anders Krogh and Lukas Käll for providing the source code of the transmembrane topology predictors and to Dmitrij Frishman and Andrei A. Mironov for early discussions. Financial support was provided by the Netherlands Bioinformatics Centre, BioRange Bioinformatics research programme (SP 3.2.2).

CHAPTER 4

Secondary structure-guided multiple sequence alignment

Submitted for publications as:

Pirovano, W., Simossis, V.A., and Heringa, J. (2009).
Secondary structure-guided multiple sequence alignment.

Abstract

Background: Current homology detection strategies use profiles and predicted secondary structure information to improve pairwise alignment quality of distantly related sequences. Given that modern secondary structure prediction techniques approach 80% in prediction accuracy, the evolutionary advantage of using secondary structure information now outweighs the chance of misprediction. Nonetheless the use of predictions in multiple sequence alignment methods is limited.
Results: We describe new ways of integrating predicted secondary structure information into the scoring scheme of our global multiple alignment program PRALINE. Our approach combines secondary structure-specific scoring matrices with appropriate gap penalty settings. Outcomes are compared to several state-of-the-art alignment methods on two alignment benchmarks and show that the inclusion of the predicted structural knowledge leads to significantly better alignments. We also investigate the independent contributions of the tailored secondary structure scheme and the specific gap penalties. Finally we describe the added value of secondary structure prediction in the context of our previously developed alignment protocol for transmembrane proteins. As a result we constructed a toolbox, called PREMIUM-Praline, that integrates homology-extended alignment with secondary and transmembrane structure information.
Conclusion: The integration of secondary structure information within the context of multiple alignment significantly improves the overall alignment quality. Secondary structure-aware alignment methods, such as the PREMIUM-Praline strategy presented here, clearly outperform methods that are guided by solely primary sequence information. The PREMIUM-Praline webserver is accessible at `www.ibi.vu.nl/programs/pralinewww`.

4.1 Introduction

In recent years, the detection of homologies between distant sequences has been significantly improved through profile-profile local alignment (Capriotti et al., 2004; Edgar and Sjölander, 2004; Ginalski et al., 2003; Jaroszewski et al., 2000; Rychlewski et al., 2000; Sadreyev and Grishin, 2003; Söding, 2005; Tomii and Akiyama, 2004; von Ohsen et al., 2004; Wang and Dunbrack, 2004; Yona and Levitt, 2002; Madera, 2008). In these approaches, single sequence input is enriched with homologous position-specific information. Some of these strategies have also incorporated structural information into their profile-profile scoring schemes to further improve the detection of distant homologies. The reason for the reported success of this incorporation is that the level of evolutionary conservation of structure is higher than that of sequence, such that the structural information can successfully anchor the alignment of distantly related sequences. Although 'true' structure information is limited, modern prediction methods give reliable predictions and are commonly invoked. A very popular tool is PSIPRED (Jones, 1999), but also other accurate methods, such SSPRO (Pollastri

et al., 2002) and PORTER (Pollastri and McLysaght, 2005), are commonly used (for a review see Pirovano and Heringa, 2009).

Recently some multiple alignment strategies have been implemented that make use of homologous sequence retrieval and predicted secondary structure (Zhou and Zhou, 2005; Pei and Grishin, 2007), originally proposed by Simossis and Heringa (2005) and Heringa (1999) The method Spem (Zhou and Zhou, 2005) employs PSI-BLAST results combined with secondary structure-dependent gap penalties. The PROMALS method (Pei and Grishin, 2007) uses PSI-BLAST profiles and predicted secondary structure to calculate probabilistic consistency scores as proposed by Do et al. (2005). In this paper we aim to integrate secondary structure information using the Lüthy secondary structure-specific substitution matrices (Lüthy et al., 1991) as originally proposed by Heringa (1999). We fully implement a profile-scoring scheme for multiple sequence alignment, following (Heringa, 1999, 2000a), in which at each alignment stage a specific matrix is chosen depending on the consistency of the predictions within the alignment profiles (see Figure 4.1). The main strength of the method resides in the 'soft manner' the predicted data is integrated. Additional improvements are then gained from optimizing the gap penalty settings. A similar protocol was developed for transmembrane proteins and showed clear benefits for the alignment quality (Pirovano et al., 2008b). We also explore the combined effect of predicted secondary and transmembrane structure information on applicable alignment cases.

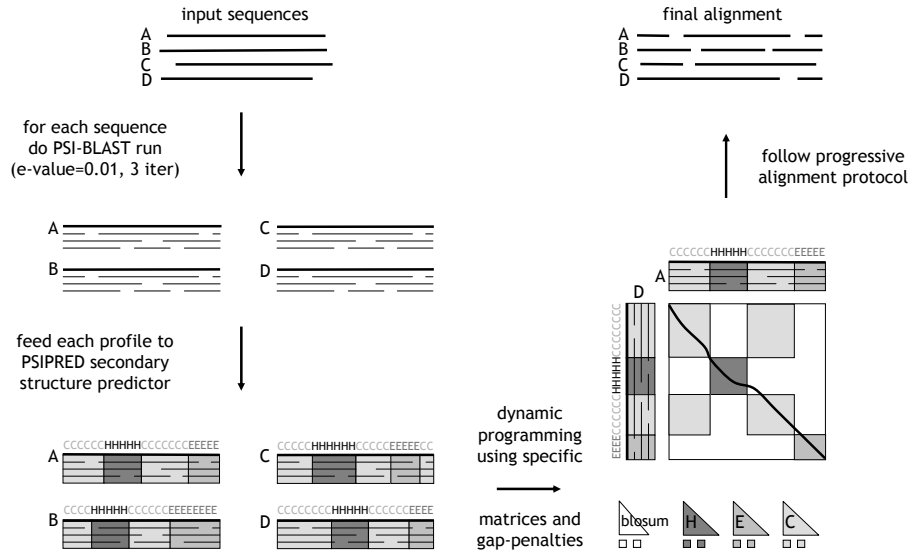


Figure 4.1: Overview of the secondary structure-guided multiple alignment strategy.

For this study we compared our strategy to several state-of-the-art methods using two multiple alignment benchmarks: BAliBASE3 (Thompson et al., 2005) and SABmark1.65 (Walle et al., 2005). We show that in many cases our algorithm outperforms

Method	Average (204)		RV11 (38)		RV12 (44)		RV20 (41)	
	TC	SP	TC	SP	TC	SP	TC	SP
ClustalW2	0.523	0.830	0.418	0.663	0.790	0.903	0.450	0.924
Muscle	0.586	**0.874**	0.541	0.743	0.826	0.931	0.486	0.941
MAFFT	0.566	0.857	0.431	0.678	0.782	0.910	0.524	0.942
Spem	0.517	0.867	0.451	0.745	0.751	0.896	0.421	0.927
$\text{PRALINE}_{\text{PSI}}$	0.556	0.839	0.479	0.711	0.800	0.920	0.543	0.942
$\text{PRALINE}_{\text{PSI + SS}}$	0.587	0.857	0.550	0.749	0.825	0.931	0.558	0.945
$\text{PRALINE}_{\text{PSI + SS gap-opt}}$	**0.604**	0.866	**0.582**	**0.769**	**0.827**	**0.934**	**0.582**	**0.951**

Table 4.1: Comparison between PREMIUM-Praline and four widely-used multiple alignment methods on the BAliBASE3 alignment benchmark sets RV11-RV50. Weighted average and individual TC (column) and SP (sum-of-pairs) scores are given. Between brackets the number of alignments for each set is given.

others in terms of alignment accuracy, which can be attributed to the inclusion of secondary structure information. The results implicate that the construction of multiple sequence alignments can directly benefit from a proper integration of predicted structural features: algorithms that do take into account predicted structural knowledge yield overall better results than those that do not use this information.

4.2 Results

4.2.1 Performance on BAliBASE

The BAliBASE benchmark comprises 217 alignments which are grouped into 6 distinct alignment cases (RV11-RV50). Findings are displayed in Table 4.1. First we compared the performance of $\text{PRALINE}_{\text{PSI}}$ (homology-extended alignment) with that of the same scheme with added secondary structure awareness ($\text{PRALINE}_{\text{PSI + SS}}$). The gap-open and gap-extension penalty for $\text{PRALINE}_{\text{PSI}}$ were set to 12 and 1, respectively, according to (Simossis and Heringa, 2005). To directly display the benefits of the secondary structure guided strategy, we applied the same gap penalties to identified helix, strand, and coil regions in the $\text{PRALINE}_{\text{PSI + SS}}$ strategy. The use of the alternative structure-specific scoring scheme leads to an average alignment accuracy of +3.1 and +1.8 percentage points for the TC- and SP score, respectively.

Since it is generally assumed that optimal gap penalty settings differ between structural elements (*i.e.* the insertion of a gap in a coil region should be less penalized than in a helical region), we have set-up an optimization scheme for secondary structure-specific gap penalties. Helix and strand gap-open penalties were varied from 12.0 to 18.0 in steps of 1.0, coil gap-open penalties from 4.0 to 18.0. For inconsistently predicted regions (where the BLOSUM62 matrix is used) standard gap-open penalties were varied from 12.0 to 15.0. The gap-extension penalty was kept at a fixed value of 1.0. From Figure 4.2 it can be clearly observed that low gap-open penalties in coil

RV30 (30)		RV40 (36)		RV50 (15)	
TC	**SP**	**TC**	**SP**	**TC**	**SP**
0.483	0.818	0.469	0.833	0.417	0.797
0.538	0.870	**0.537**	**0.872**	0.483	0.872
0.557	0.868	0.521	0.869	**0.520**	0.867
0.538	**0.896**	0.434	0.864	0.421	**0.874**
0.565	0.843	0.419	0.776	0.381	0.787
0.601	0.867	0.434	0.792	0.406	0.810
0.624	0.881	0.439	0.797	0.415	0.810

regions have a positive effect on the alignment. In other regions different gap penalty settings have less influence on the alignment quality and trends are not as clear. The same conclusions can be drawn from Figure 4.3, where the gap settings are analyzed using a principal component analysis (PCA), clearly pointing out the significance of the coil gap-open penalty value.

The optimal gap-open penalties are 12.0, 13.0, 13.0, and 6.0 for standard, helical, strand, and coil regions, respectively. These settings define the final PRALINE strategy ($\text{PRALINE}_{\text{PSI + SS gap-opt}}$). Results are compared to other state-of-the-art methods: our PRALINE method yields the highest TC score on average and on most individual sets (RV11-RV30; Table 4.1). Muscle achieves the best SP score, even if for half of the individual sets PRALINE shows a better performance (RV11-RV20). The RV40 and RV50 sets, which are characterized by sequences containing large N/C-terminal extensions and internal insertions, appear to be rather difficult to align. On set RV40 Muscle is best while on set RV50 non of the competitors clearly wins. Note that for 13 cases in set RV40 the method Spem was unable to produce alignments. Methods are therefore compared over a subset of 36 alignments ($49-13$).

Summarizing Table 4.1 we observe that on average PRALINE performs best, Muscle is runner-up while ClustalW2 performs worst. The method Spem, which also makes use of predicted secondary structure information, gives scores comparable to MAFFT. In contrast, it should be noticed that for each set the performance of the secondary structure-extended protocol ($\text{PRALINE}_{\text{PSI + SS}}$) achieves better results than the original $\text{PRALINE}_{\text{PSI}}$ strategy. For each set the alignment quality is further improved by using the optimal gap penalty settings ($\text{PRALINE}_{\text{PSI + SS gap-opt}}$).

Table 4.2 sheds a slightly different light on the results. Here PRALINE and its contenders are compared in terms of cases won by each individual method. PRALINE makes the best alignment in 77 cases though Spem is second best (69 cases won). Interestingly, these findings indicate that in most cases the secondary structure-guided multiple alignment tools construct better alignments (146 cases) even though a sig-

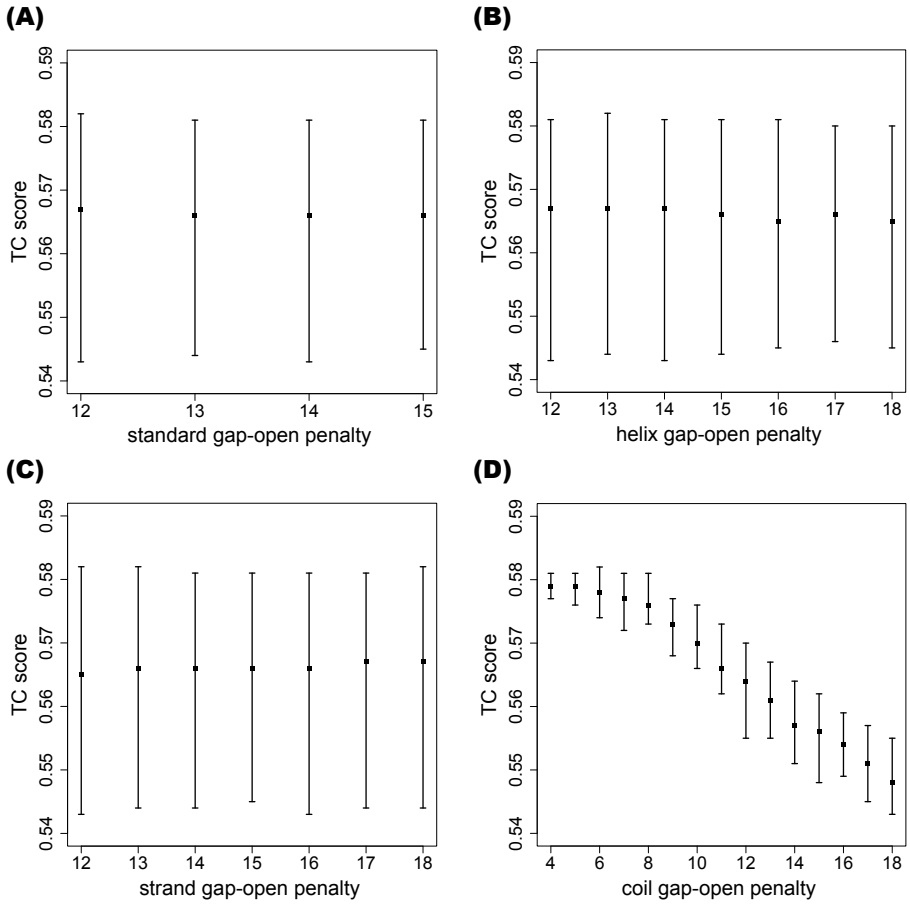


Figure 4.2: Gap optimisation on the BAliBASE3 alignment benchmark set. Alignment accuracy in terms of the 'TC score' is shown for different (**A**) standard, (**B**) helix, (**C**) strand, and (**D**) coil gap-open penalties.

nificant part is won by the other three methods (107 cases).

4.2.2 Performance on SABmark

The SABmark benchmark contains around three times more alignments than BAliBASE: 627 in total which are subdivided into a Superfamilies (422) and a Twilight Zone (205) alignment set. For both sets the secondary structure-aware PRALINE strategies outperform the original $\text{PRALINE}_{\text{PSI}}$ protocol. Moreover the alignment quality benefits from the adjusted gap penalties. These findings are fully in line with our previous results on BAliBASE, even though SABmark was not used for train-

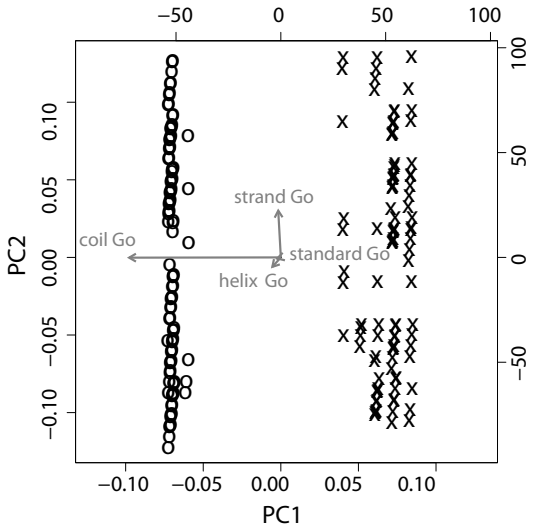


Figure 4.3: Principal component analysis displaying the importance of appropriate structure-specific gap penalty settings, especially for coil regions. The input consists of all average BAliBASE3 'TC scores' obtained using the different gap-penalty combinations (a total of 2940 scores). 'O' represent the 100 best 'TC scores' while 'X' represent the 100 worst 'TC scores'.

Method	**Wins**	**%**
ClustalW2	32	15.7
Muscle	43	21.1
MAFFT	32	15.7
Spem	69	33.8
$\text{PRALINE}_{\text{PSI + SS gap-opt}}$	**77**	**37.7**

Table 4.2: Number of wins on BAliBASE3. The total number of alignments for which a method yields the best 'TC score' is given. In the second column also the corresponding percentages are provided. Note that percentages do not sum up to 100% due to cases where different methods give identical alignments.

ing the four gap-open parameters. Strikingly, Spem clearly performs best on this benchmark and PRALINE is runner up before Muscle, MAFFT, and ClustalW2. Results are summarized in Table 4.3. The great potential of secondary structure-guided alignment methods is also strongly underscored by comparing the number of wins obtained by each method (see Table 4.4: in the majority of cases Spem and PRALINE

Method	Average (627)		sup (422)		twi (205)	
	f_D	f_M	f_D	f_M	f_D	f_M
ClustalW2	41.876	30.797	51.079	38.321	22.931	15.308
Muscle	43.775	32.502	53.157	40.120	24.462	16.820
MAFFT	42.569	32.034	52.124	39.757	22.899	16.135
Spem	**60.434**	**44.581**	**68.301**	**51.136**	**44.238**	**31.086**
$\text{PRALINE}_{\text{PSI}}$	47.559	34.941	56.804	42.538	28.527	19.303
$\text{PRALINE}_{\text{PSI + SS}}$	50.425	36.988	59.676	44.609	31.382	21.300
$\text{PRALINE}_{\text{PSI + SS gap-opt}}$	51.484	37.753	60.640	45.304	32.637	22.209

Table 4.3: Comparison between PREMIUM-Praline and competitors on the SABmark alignment sets Superfamilies (sup) and Twilight Zone (twi). Individual f_D and f_M scores are given. Between brackets the number of alignments for each set is given.

Method	Wins	%
ClustalW2	53	8.5
Muscle	59	9.4
MAFFT	38	6.1
Spem	**443**	**70.7**
$\text{PRALINE}_{\text{PSI + SS gap-opt}}$	143	22.8

Table 4.4: Number of wins on SABmark. The total number of alignments for which a method yields the best f_D score is given. In the second column also the corresponding percentages are provided. Note that numbers do not sum up due to cases where different methods give identical alignments.

construct the best alignment (586 wins versus 150 for standard methods).

4.2.3 Alignment of transmembrane proteins

Finally, we evaluated the combined effect of integrating both secondary and transmembrane structure prediction into the protocol. As a reference we used the BAliBASE 'ref7' set, which consists of eight aligned transmembrane protein families. Since Spem was unable to correctly complete one of the alignments, comparisons are given for seven families only. From Table 4.5 it can be noticed that the PRALINE homology-extended protocol ($\text{PRALINE}_{\text{PSI}}$) can be improved by including secondary structure information ($\text{PRALINE}_{\text{PSI + SS}}$), though particular attention has to be paid to gap penalty settings. In fact the optimal gap settings found for soluble proteins ($\text{PRALINE}_{\text{PSI + SS gap-opt}}$) do not apply to soluble regions of transmembrane proteins, *i.e.* lead to decreased performance, and have to be adjusted accordingly.

Method	TC	SP
ClustalW2	0.339	0.769
Muscle	0.361	0.828
MAFFT	0.266	0.796
Spem	0.394	**0.855**
PRALINE^{TM}*	**0.436**	0.837
PRALINE$_{PSI}$	0.333	0.766
PRALINE$_{PSI + SS}$	0.371	0.770
PRALINE$_{PSI + SS\ gap\text{-}opt}$	0.313	0.773
PRALINE$_{PSI + SS + TM\ gap\text{-}opt}$	0.406	0.833

* *cf.* Pirovano et al. (2008b).

Table 4.5: Comparison between different PRALINE strategies displays the benefits of including predicted structural features. Results are shown for the BAliBASE 'ref7' transmembrane alignment set by means of the TC (column) and SP (sum-of-pairs) scores and compared to other alignment methods.

As a next step we also included predicted transmembrane information (applying a standard gap-open penalty of 15.0 and transmembrane gap-open penalty of 16.5 following Pirovano *et al.* (2008b). The combination of both types of predicted information leads to a rather significant increase in performance. Regarding the gap penalties, our results indicate that highest average accuracy is achieved using gap-open penalties of 15.0, 14.0, 13.0, 12.0, and 16.5 for standard, helical, strand, coil, and transmembrane regions, respectively (PRALINE$_{PSI + SS + TM\ gap\text{-}opt}$). Our previously published PRALINE^{TM} tool obtains a somewhat better performance regarding the 'TC score', but this can mainly be attributed to the fact that the new strategy applies a general PSI-BLAST e-value of 0.01 instead of the so-called 'pre-profiles' from pairwise sequence comparison within the input sequences. The use of a general e-value cut-off makes the PREMIUM-Praline alignment procedure straightforward and transparent to use for any sequence set, although for known membrane protein sequence sets the PRALINE^{TM} strategy is recommended. Spem performs somewhat better regarding the 'SP score' and, even if the method could not align all input sequence sets, the added value of predicted secondary structure clearly extends to transmembrane proteins.

4.3 Discussion and conclusion

In this paper we introduced an integrative multiple alignment toolbox, called PREMIUM-Praline, which uses predicted secondary structure information to guide the alignment process. The information is integrated in a soft manner, meaning that different structural regions are aligned using specific evolutionary schemes, rather than imposing the self-alignment of secondary structure elements. Impressively, for all benchmark sets used in this study, the addition of predicted information has a

clearly positive effect on the alignment accuracy. On all sets the quality is further enhanced using gap optimized penalty settings which are tailored to the structural regions.

These findings are fully in line with our previous results for transmembrane proteins (Pirovano et al., 2008b). Here we have further extended the transmembrane protocol by employing predicted secondary structure to the soluble parts of these proteins. Also for these cases the advantage of combined predictions is notable compared to other recent state-of-the-art methods. We also found that optimal gap penalties for soluble parts of transmembrane proteins differ from those used for soluble proteins in general. This indicates that different evolutionary schemes apply to membrane-bound versus soluble proteins.

Overall PRALINE performs best or second-best: on average our method scores best on the BAliBASE3 benchmark set, but is runner-up on SABmark. Over the latter set most accurate alignments are produced by Spem, which proposes an alternative way of integrating predicted secondary structure data. Unfortunately Spem fails to produce an alignment in a significant number of cases. There are two main differences between the alignment strategy implemented in PRALINE and Spem. PRALINE adopts a more sophisticated inclusion of structural features by employing *ad hoc* residue exchange weights in combination with optimal gap penalty-settings whereas Spem uses a simpler procedure that incorporates structure-based gap penalties. On the other hand, the profile-profile comparison methodology of PRALINE is more straightforward than the SP^2 pairwise alignment refinement step used by Spem. It might be that the different profile comparison strategies have a pronounced influence on the alignment quality, such that investigating alternative profile-profile scoring methods may further improve our method.

In summary, this study evidently underscores that future developments in the field of protein sequence alignment should take predicted structural information into account. It can be argued that a main pitfall of these methods is that they are relatively time-consuming due to retrieval of homologous sequences by PSI-BLAST (which is essential for high-quality secondary structure predictions). However secondary structure-guided alignment gives undisputedly better results compared to standard protocols and for many biological applications the improved alignment quality outweighs the temporal disadvantage.

4.4 Methods

The PREMIUM-Praline progressive alignment method presented here is based on the previously published $\text{PRALINE}_{\text{PSI}}$ algorithm (Simossis et al., 2005). In this homology-extended strategy a PSI-BLAST search (Altschul et al., 1997) is invoked for each sequence in the given set to collect potential homologues that score above a predetermined e-value cut-off. This extended collection of homologues for each sequence is represented as a homology-extended profile and used as the starting point

for the progressive routine, instead of the individual query sequences in the set.

The addition to PRALINE$_{\mathrm{PSI}}$ tool we present here is a profile-scoring scheme that integrates secondary structure-specific information in the form of secondary structure-specific substitution scores from the Lüthy series of matrices (Lüthy et al., 1991). In this regard, the homology-extended profiles are also used as input for predicting the secondary structure for each of the query sequences. Then, the secondary structure information is assigned to all the hits in the profile generated from database searching. This way, each homologue in the homology-extended profile is assigned the secondary structure of the query (top) sequence. This generalization of the local structure of the homology-extended profile sequences is necessary because re-running predictions for all of them would be computationally prohibitive and biologically uncertain given that homologous fragments detected by PSI-BLAST can be relatively short. An overview of the final PREMIUM-Praline strategy is given in Figure 4.1.

We also included an advanced option for alignment of transmembrane proteins which extends our PRALINE$^{\mathrm{TM}}$ method (Pirovano et al., 2008b). In this scenario, similarly to the secondary structure protocol, the homology-extended profiles are fed to a transmembrane topology predictor and integrated in our transmembrane-aware profile-scoring scheme. The transmembrane structure-specific substitution scores are assigned using the PHAT substitution matrix (Ng et al., 2000).

4.4.1 Profile scoring

In the standard PRALINE scoring scheme, a pair of profile columns x and y is scored as follows:

$$S(a,b) = \sum_{i}^{20} \sum_{j}^{20} \alpha_i \beta_j M(i,j) \tag{4.1}$$

where α_i and β_j are the frequencies with which residues i and j appear in columns x and y, respectively, and $M(i,j)$ is the exchange weight for residues i and j provided by the selected substitution matrix M (*e.g.* the PRALINE default is BLOSUM62).

In the new scoring scheme we propose, when all residues being compared from each profile column are assigned the same secondary structure type (H, E or C), the corresponding secondary structure-specific 'Lüthy' substitution matrix (Lüthy et al., 1991) is used. Otherwise, the BLOSUM62 matrix (or any other chosen matrix) is used. Analogously, if the transmembrane alignment option is set and two columns are consistently predicted to be transmembrane, the PHAT matrix (Ng et al., 2000) is employed. Predicted secondary structure information is only employed for non-TM regions, *i.e.* the TM predictions override the SS predictions.

In this way the predicted structural information is integrated in a 'soft' manner: the alignment of structural elements is guided by a mixed evolutionary scheme (the different matrices) rather than imposing self-pairing of the elements. All substitution matrices were normalized using their diagonal elements as described in (Abagyan and Batalov, 1997).

4.4.2 Secondary and transmembrane structure predictors

By default the PREMIUM-Praline algorithm invokes PSIPRED v2.6 (Jones, 1999) to predict the three-state secondary structure elements (H, E, and C). Alternatively the user may specify several other predictors: SSpro v4.0 (Baldi et al., 1999), PORTER (Pollastri and McLysaght, 2005), and YASPIN (Lin et al., 2005). Additionally, a DSSP option (Kabsch and Sander, 1983) can be set to determine the 'true' secondary structures for sequences that have solved three-dimensional structure co-ordinates in the PDB (Berman et al., 2000). In case of transmembrane proteins, three different topology predictors can be selected: the method of choice is PolyPhobius (Käll et al., 2005), but also HMMTOP v2.1 (Tusnády and Simon, 2001) and TMHMM (Krogh et al., 2001) have been integrated into the protocol. All methods (except for TMHMM) can use PSI-BLAST profile information for their predictions. Therefore we used the profiles generated by the strategy of $\text{PRALINE}_{\text{PSI}}$Simossis et al. (2005) as input for the predictions. For all runs the PSI-BLAST program was invoked using 3 iterations and an e-value cut-off of 10^{-2} on the non-redundant database (NR update 05/2008).

4.4.3 Alignment quality assessment

We used three different alignment benchmark databases as our reference. Optimization of the parameters was performed on BAliBASE3 reference sets 'RV11 to RV50' (Thompson et al., 2005), while validation was done (also) on SABmark1.65 (Walle et al., 2005) and the BAliBASE3 transmembrane reference set ('ref7'). 'RV11 to RV50' correspond to the complete set of soluble proteins (but without the transmembrane set 'ref7') and covers roughly 25% of all alignments considered in this study. The alignment quality was assessed using the individual 'testing' programs provided with each benchmark set. For BAliBASE the 'TC score' measures the number of correctly aligned columns in the test alignment to the number of columns in the reference alignment while the 'SP score' measures the number of correctly aligned residue pairs in the test alignment to the number of aligned pairs in the reference alignment. Hence the TC score is the stricter measure. SABmark is tested by dividing the number of correctly aligned residues in the test alignment by total number of pairs of either the reference alignment ('developer (f_D) score', equivalent to the SP score) or the test alignment (the 'modeler (f_M) score').

4.4.4 Competitors

The performance of the PREMIUM-Praline strategy was compared to four other current state-of-the-art multiple alignment methods: ClustalW2 (Larkin et al., 2007), MUSCLE v3.7 (Edgar, 2004a), MAFFT v6.240 (Katoh et al., 2005) and Spem (Zhou and Zhou, 2005). Programs were run using default parameter settings.

4.4.5 System

All methods were run locally on the IBIVU server (64 Intel Xeon 3.0GHz). The PREMIUM-Praline tool is written in the C programming language and the source code can be made available from the authors upon request.

4.5 Authors‘ contributions

WP and VAS performed the research. All authors significantly contributed to the design of this study and the analysis of the results. All authors were involved in writing the manuscripts and read and approved its final version.

4.6 Acknowledgements

The authors wish to thank K. Anton Feenstra and Bart P.P. van Houte for their contribution to the software development and precious advice. We are also grateful to Bernd W. Brandt for his help in method evaluation. Financial support was provided by the Netherlands Bioinformatics Centre, BioRange Bioinformatics research programme SP 3.2.2.

CHAPTER 5

Sequence comparison by sequence harmony identifies subtype-specific functional sites

Published as:

Pirovano, W., Feenstra, K.A., and Heringa, J. (2006).
Sequence comparison by sequence harmony identifies subtype-specific functional sites.
Nucleic Acids Res, 34(22):6540–6548.

Abstract

Multiple sequence alignments are often used to reveal functionally important residues within a protein family. They can be particularly useful for the identification of key residues that determine functional differences between protein subfamilies. We present a new entropy-based method, Sequence Harmony (SH) that accurately detects subfamily-specific positions from a multiple sequence alignment. The SH algorithm implements a novel formula, able to score compositional differences between subfamilies, without imposing conservation, in a simple manner on an intuitive scale. We compare our method with the most important published methods, *i.e.* AMAS, TreeDet and SDP-pred, using three well-studied protein families: the receptor-binding domain (MH2) of the Smad family of transcription factors, the Ras-superfamily of small GTPases and the MIP-family of integral membrane transporters. We demonstrate that SH accurately selects known functional sites with higher coverage than the other methods for these test-cases. This shows that compositional differences between protein subfamilies provide sufficient basis for identification of functional sites. In addition, SH selects a number of sites of unknown function that could be interesting candidates for further experimental investigation.

5.1 Introduction

In the quest for knowledge about protein function, understanding differences between protein families is essential. It is therefore not surprising that a large number of methods have already been introduced for the positional comparison of amino acid compositions between different protein families or subtypes (Livingstone and Barton, 1996; Lichtarge et al., 1996; Kuipers et al., 1997; Hannenhalli and Russell, 2000; Mirny and Gelfand, 2002; del Sol Mesa et al., 2003; Kalinina et al., 2004; Donald and Shakhnovich, 2005; Ye et al., 2006; for a review see Whisstock and Lesk, 2003). Though these methods have contributed greatly to the understanding of the relation between sequence and function (Obenauer et al., 2006), coverage of known sites of functional differences has been limited.

An early method called AMAS by Livingstone and Barton (1996) analyses conservation patterns using a number of physiochemical properties of amino acids. The method assigns sites to either of three classes: globally conserved, conserved in subfamilies, or not conserved, based on arbitrarily set conservation threshold values. This method was intended primarily to 'allow the residue-specific similarities and differences in physiochemical properties between groups of sequences to be identified quickly' (Livingstone and Barton, 1996).

Several methods have been based on evolutionary trace analysis (Lichtarge et al., 1996) and rely on residue conservation within subfamilies to construct subfamily-specific consensus sequences. These sequences are then aligned to reveal the variation between the different subfamilies. Kuipers et al. (1997) relaxed the requirement of intra-group conservation and allowed the selection of 'residues that are conserved in

one class of proteins with a certain function but are different in other classes'.

del Sol Mesa et al. (2003) introduced a method called TreeDet that uses mutational behaviour analysis of so-called 'tree-determinant' residues. The method uses an internal algorithm for unsupervised grouping of the input sequences. It then selects residues that follow mutation patterns similar to that of the overall phylogeny. For each alignment position, this is measured using the correlation coefficient between the substitution matrix derived for the position considered and that for the whole protein. Residues selected tend to be conserved within each subfamily but different between them. The method aims to find 'the most appropriate way of dividing a protein family into subfamilies in order to associate the tree-determinants with sites, which are likely to be responsible for functional differences between these subfamilies'.

The method SDP-pred by Kalinina et al. (2004) selects residues that 'are well conserved within specificity groups but differ between these groups'. The method is based on mutual information analysis and complex statistical treatment. It extends an earlier method (Mirny and Gelfand, 2002) by using residue frequencies smoothed and weighted by the BLOSUM average substitution scores. Significance is estimated by Z-scores of expected mutual information obtained from column shuffling. Subsequently, an appropriate Z-score threshold for selection of high-ranking sites is determined using the Bernoulli estimator. Recently, Donald and Shakhnovich (Donald and Shakhnovich, 2005) added an automated procedure for functional grouping.

The above methods focus on sites that are conserved in one or both groups and subsequently select those sites that are different between these groups (Whisstock and Lesk, 2003). Unfortunately, this scenario excludes sites that are not highly conserved within each of the groups. This may not seem a serious problem at first hand, but let us consider an example. Take a group A comprising a protein subfamily binding a certain molecule (*e.g.* a ligand or receptor; this group represents the 'binders') and group B comprising proteins that do not (the 'non-binders'). Certainly, one can expect sites that are crucial for binding to be conserved in group A (the 'binders'). However, for group B to avoid binding, the corresponding site would need to avoid the conserved residue of group A. It seems therefore imprudent to expect conservation throughout group B (the 'non-binders') as well (Kuipers et al., 1997). Moreover, if group A contains binders to different (but related) molecules (ligands or receptors), even the requirement of conservation in group A may not apply.

To address these restrictions here, we introduce an alternative similarity measure for comparing groups of sequences within a multiple sequence alignment, which we name Sequence Harmony (SH). The SH measure is derived from Shannon's general information entropy (Shannon, 1948) as applied to biomolecular sequences by Shenkin et al. (1991). We will show that SH has well-defined properties and is easily calculated. The values obtained fall within a convenient fixed interval and correspond intuitively to differences in the amino acid compositions as observed in the alignment. Moreover, no additional residue substitution matrices, calculations of mutual information or computations of statistical significance (Mirny and Gelfand, 2002; Kalinina et al., 2004).

We evaluate the SH method using four benchmark sets comprising experimental

data on specificity-switching residues and compare its performance against three other state-of-the-art methods. From this, we obtain new insights about the principles that govern the accurate detection of functional sites. We will show that SH is able to identify sites associated with subfamily specificity systematically and with relatively few errors.

5.2 Theory

5.2.1 Comparing sequences by relative entropy

Relative entropy (rE) is commonly used in sequence comparison to quantify the degree of conservation (Heger et al., 2004; Mihalek et al., 2004). It is derived from Shannons general information entropy (Shannon, 1948) as applied to biological sequences by Shenkin et al. (1991):

$$rE_i^{A/B} = \sum_x p_{i,x}^A \log \frac{p_{i,x}^A}{p_{i,x}^B} \tag{5.1}$$

where $p_{i,x}^A$ and $p_{i,x}^B$ are the observed probabilities of amino acid type x at a position i in the alignment of groups A and B, respectively. Relative entropy measures the difference in information content between both distributions of amino acid types. Interestingly, for sites to be maximally different between the two groups, amino acid types in one group should be absent in the other or vice versa. This leads to a degenerate result (singularity) whenever the entropy function of Equation 5.1 is used. Inclusion of so-called 'pseudo-counts' (Henikoff and Henikoff, 1996) solves the degeneracy but not the unbounded, asymptotic behaviour of Equation 5.1. Also taking, for instance, the relative entropy with respect to both groups, as $rE_i^{A/B} = \sum_x p_{i,x}^A \log \left(p_{i,x}^A / p_{i,x}^B\right)$, does not solve the unbounded behaviour. Moreover, the upper limit depends on the ratio of the number of sequences in both groups.

5.2.2 Calculating differences by SH

We address the degeneracy of Equation 5.1 by defining SH as

$$SH_i^{A/B} = \sum_x p_{i,x}^A \log \frac{p_{i,x}^A}{p_{i,x}^A + p_{i,x}^B} \tag{5.2}$$

This can be viewed as the relative entropy of group A relative to the sum of the probabilities of both groups ($pA+pB$). As a consequence of the separate weighting of both groups, we eliminate the dependence on relative group sizes. Since, in general, $SH^{A/B} \neq SH^{B/A}$, we remedy this non-commutativeness by taking the average. Using Equation 5.2 this leads to:

$$\begin{aligned} SH_i &= \frac{1}{2}\left(SH_i^{A/B} + SH_i^{B/A}\right) \\ &= \frac{1}{2}\left(\sum_x p_{i,x}^A \log p_{i,x}^A + \sum_x p_{i,x}^B \log p_{i,x}^B - \sum_x (p_{i,x}^A + p_{i,x}^B)\log(p_{i,x}^A + p_{i,x}^B)\right) \end{aligned} \tag{5.3}$$

Using Shannons information entropy:

$$S_i = -\sum_x p_{i,x} \log p_{i,x} \tag{5.4}$$

and writing

$$S_i^{A+B} = -\sum_x \left(p_{i,x}^A + p_{i,x}^B\right) \log \left(p_{i,x}^A + p_{i,x}^B\right) \tag{5.5}$$

as Shannons entropy weighting groups A and B equally, we can rewrite Equation 5.3 as

$$SH_i = \frac{1}{2}\left(S_i^{A+B} - S_i^A - S_i^B\right) \tag{5.6}$$

In other words, SH can also be conveniently expressed as a simple linear combination of entropies. The SH function juxtaposes relative entropy since it becomes zero for maximally different sites and one for sites with identical distributions. We therefore coined the phrase *Sequence Harmony* (SH) as it indicates to what extent amino acid compositions between two groups of sequences are in harmony.

To illustrate how SH works in practice, we present a hypothetical alignment in Table 5.1. Positions 1, 2 and 3 all have a SH value of zero. In fact, the formula becomes zero any time the amino acid composition in one group is non-overlapping with that in the other group, regardless of conservation. At position 4, the amino acid of group A (Ala) also occurs once in group B, yielding the lowest possible non-zero score for this number of sequences. For positions 5 and 6, there is an increasing overlap and hence growing SH values. Position 8 is conserved overall in the whole family and is therefore maximally harmonious with an SH value of one. Note that unconserved sites have SH = 1 whenever the two groups have identical compositions. This is illustrated in position 7, where equal proportions of R and K are shown.

The SH measure is implemented in a simple online server for calculating SH from an alignment. It can be accessed at `www.ibi.vu.nl/programs/seqharmwww`.

5.2.3 Ranking identically scoring SH sites

The example of Table 5.1 shows that sites with different compositions can have identical SH values, which therefore cannot be ranked. This is particularly important for sites with SH = 0, as these are potentially discriminative for function. To address this issue, we derived a simple but effective ranking for identically scoring sites based on

	Alignment position							
	1	2	3	4	5	6	7	8
Group A								
seq1	R	E	L	A	A	A	K	K
seq2	R	E	L	A	A	F	K	K
seq3	R	E	A	A	A	Y	R	K
seq4	R	E	A	A	A	F	R	K
Group B								
seq1	H	N	V	A	A	Y	R	K
seq2	H	N	V	F	F	Y	R	K
seq3	H	N	F	F	A	F	K	K
seq4	H	S	F	F	F	Y	R	K
seq5	H	S	M	F	F	F	K	K
seq6	H	T	M	F	F	Y	K	K
SH	0.00	0.00	0.00	0.35	0.54	0.79	1.00	1.00

Table 5.1: Hypothetical alignment of two subfamilies A (4 sequences) and B (6 sequences. For each position in the alignment the SH score between the subfamilies is calculated.

the distribution of low SH values (SH ≤ 0.2) over the alignment positions. Groups of sequentially adjacent low-harmony sites can have one intermediate high-harmony site in between, *i.e.* the sequence distance is two or less. The ranking is first on increasing SH (low harmony first), and then on decreasing group size (larger sequential groups first). Finally, sites that have equal SH scores and group sizes are ranked on the total entropy of the sites in both subfamilies. This scenario is implemented in the SH method which we will label 'SH'.

We further explored the performance of an even simpler ranking on just SH values and entropy of the site, *i.e.* without ranking on sequential groups. This approach we will refer to as 'SH/E'.

5.2.4 Benchmarking

We compared predictions by SH with those obtained from three other methods over several protein families. For this purpose, we have used the online server of AMAS (`http://barton.ebi.ac.uk/servers/amas_server.html`) (Livingstone and Barton, 1996), the 'mutational behaviour' method available from the TreeDet server, which we will refer to as TreeDet/MB (`http://somosierra.cnb.uam.es/Servers/treedetv2/`) (del Sol Mesa et al., 2003) and the SDP-pred server (`http://math.genebee.msu.ru/~psn/`) (Mirny and Gelfand, 2002; Kalinina et al., 2004). The recent method by Donald and Shakhnovich (Donald and Shakhnovich, 2005) (see Section 5.1) could not be evaluated in this study

due to lack of an online server.

The methods were evaluated over a number of test-sets (see below) using receiver-operator characteristics (ROC) plots of coverage (% recovered known functional sites) versus error (% wrong predictions). These were constructed from ranked lists of sites for SH and for SDP-pred. For AMAS, the conservation threshold was varied from 0 to 10 (see Section 5.1). For TreeDet/MB, the cut-off value was varied from 0 to 1.0. In both cases, small steps were taken so that the addition of single sites to the selection could be observed.

5.3 Datasets for validation

We have selected three relevant protein families, for which experimental evidence is available on subfamily-specific sites. These include families that have been used for construction, testing and/or validation of the other prediction methods and one family for which we have assembled an extensive dataset.

In many cases, the available experimental evidence concerns the exchange of a sequence segment, *e.g.* a loop or a helix. Positions within such segments that are conserved over both subfamilies were not regarded as subfamily specific. Although such positions are likely to be important for the function of the family, they are unable to explain functional differences between subfamilies at the residue level.

5.3.1 TGF-β-associated transcription factors (Smad family)

The Smad family of transcription factors plays a crucial role in the transforming growth factor-β (TGF-β) signaling pathway. Smads are also critical for determining the specificity between alternative TGF-β pathways [for recent extensive reviews, see Feng and Derynck (2005) and Massagué et al. (2005)]. This complex signaling network is involved in the regulation of many cellular processes such as division and differentiation, motility, adhesion and programmed cell death. The TGF-β family of growth factors induces Type-I transmembrane receptors to phosphorylate and activate the receptor-regulated Smads (R-Smads) (Massagué et al., 2005). The R-Smads can be subdivided into two major groups: the AR-Smads, which are mainly induced by TGF-β-type receptors (TβR-I), and the BR-Smads, which are mainly induced by the BMP-type receptors (BMPR-I and ALK1/2). Subsequent associations among Smads are responsible for control of TGF-β target genes in the nucleus. It has been shown that most of the above interactions involve the so-called Mad Homology 2 (MH2) domain of the Smad proteins (Feng and Derynck, 2005). From an extensive literature search, we have identified 29 specific sites in the MH2 domain that are experimentally validated to be important for Smad specificity, as listed in Table 5.2.

5.3.2 Small GTPases (RAS superfamily)

Members of the Ras superfamily of GTPases are implicated in the regulation of growth, survival, differentiation and other processes in haematopoietic cells (Reuther

Position		Sec. struc.	Consensus		SH	SH group size	Other methods		
Align	Smad2		AR	BR			AMAS	TreeDet	SDPpred
2	(L263)	B1'	La	Vfm	0	1	+	−	−
3	(Q264)	B1'	Qa	Qrh	0.81	−	−	−	−
6	T267	B1'	Tm	Acen	0	2	−	−	−
8	S269	loop	CSh	Eq	0	2	−	−	−
11	A272	loop	A	Kqls	0	2	−	−	−
12	F273	loop	F	Hy	0	2	−	−	−
23	Q284	B2	Qt	N	0	1	−	−	−
33	Q294	loop	Q	Sq	0.16	4	−	−	−
34	P295	B3	P	Trl	0	4	−	0.85	−
36	L297	B3	LMi	Vi	0.11	4	−	−	−
37	T298	B3	T	Li	0	4	−	0.88	−
47	S308	L1	Sa	N	0	3	−	−	−
48	−	L1	−	Nsd	0	3	−	−	−
49	E309	L1	E	Krs	0	3	−	−	−
63	A323	H1	Ae	S	0	3	−	0.84	−
65	V325	H1	V	I	0	3	−	0.87	2.17
67	M327	H1	LMq	N	0	3	+	0.83	−
74	R334	loop	Rk	K	0.18	1	−	−	−
77	R337	B5	R	H	0	1	−	0.87	2.25
81	I341	B5	I	V	0	1	−	0.87	2.24
86	F346	B6	F	Y	0	1	−	0.87	2.14
94	A354	B7	As	S	0.18	1	−	−	−
100	P360	H2	P	R	0	1	+	0.87	2.21
104	Q364	H2	Q	Yf	0	4	−	−	−
105	R365	H2	R	Hq	0	4	−	−	−
106	Y366	H2	Y	H	0	4	−	0.86	2.02
108	W368	loop	W	F	0	4	−	0.87	2.22
118	P378	loop	P	Sp	0.16	1	−	−	−
121	N381	B9	N	S	0	1	−	0.87	−
136	A392	H3	A	Qeh	0	1	−	0.85	−
144	Q400	loop	Q	H	0	1	+	0.87	2.03
151	Q407	H4	Qr	E	0	1	−	0.83	−
154	R410	H4	R	K	0	1	−	0.87	2.28
171	R427	L3	R	H	0	1	−	0.87	2.01
174	T430	L3	T	D	0	1	−	0.87	2.08
184	L440	B11	L	Iv	0	1	−	0.84	−
187	N443	H5	N	Hn	0.16	1	−	−	−
204	S460	C-tail	Snr	Hlr	0.06	4	−	−	−
205	V461	C-tail	Ivl	N	0	4	+	0.83	−
206	R462	C-tail	Rp	P	0.17	4	−	−	−
207	C463	C-tail	C	I	0	4	+	0.86	2.30
210	M466	C-tail	MV	V	0.69	−	−	−	−

Function	Refs
SARA	(1)
SARA	(1)
SARA	(1)
? (SARA)	–
? (SARA)	–
? (SARA)	–
TβR-I	(2)
c-Ski/SnoN	(3)
c-Ski/SnoN	(3)
c-Ski/SnoN	(3)
c-Ski/SnoN	(3)
c-Ski/SnoN	(3)
c-Ski/SnoN	(3)
c-Ski/SnoN	(3)
BMPR-I/ALK1/2	(4)
BMPR-I/ALK1/2	(4)
BMPR-I/ALK1/2	(4)
? (c-Ski/SnoN)	–
not SARA (c-Ski/SnoN)	(1)
SARA/Mixer	(1,5)
SARA/Mixer	(1,5)
?	–
FAST1	(6)
Mixer/FAST1	(6,7)
Mixer/FAST1	(6,7)
SARA/Mixer/FAST1	(1,5,6,7)
SARA/Mixer/FAST1	(1,5,6)
?	–
SARA/Mixer	(1,5)
? (FAST1/Mixer)	–
? (FAST1/Mixer)	–
not receptor binding (FAST1/Mixer)	(8)
? (FAST1/Mixer)	–
TβR-I/BMPR-I/ALK1/2	(8)
TβR-I/BMPR-I/ALK1/2	(8)
?	–
? (SARA)	–
TβR-I/BMPR-I	(8)
TβR-I/BMPR-I	(8)
TβR-I/BMPR-I	(8,9)
TβR-I/BMPR-I	(8,9)
TβR-I/BMPR-I	(8)

Table 5.2: Summary of all known functional sites and sites of low SH (SH <0.2) in the MH2 domain of Smad. Sequence positions are indicated relative to the alignment (Align) as well as to Smad2, according to PDB 1KHX (Wu et al., 2000). Secondary structure elements are indicated according to Chen et al. (1998). SH scores and corresponding size of sequential groups of low SH sites are listed. In addition, predictions for the other methods using default settings are shown (see text for details). The consensus patterns for the AR-Smads and BR-Smads are shown. All amino acid types are listed in order of decreasing frequency. Those of half or less than the frequency of the dominant type are in lower case. The known functions, with corresponding reference(s) (see below for numbering), are indicated. Putative functions corresponding to structural clustering shown in Figure 5.5A are indicated in brackets (see text for more detail). (1) Wu et al., 2000 (2) Huse et al., 2001 (3) Mizuide et al., 2003 (4) Chen and Massagué, 1999 (5) Randall et al., 2002 (6) Chen et al., 1998 (7) Germain et al., 2000 (8) Lo et al., 1998 (9) Yakymovych et al., 2004

and Der, 2000). They comprise six families. Experimental evidence for functional sites is available from the literature for the Ras versus Ral families (del Sol Mesa et al., 2003; Bauer et al., 1999). This set was used by del Sol Mesa et al. (2003) for the development of the TreeDet method. In addition, we include a test set of Rab5/6-specific sites obtained from (Stenmark and Olkkonen, 2001; Stenmark et al., 1994).

5.3.3 Integral membrane transporters (MIP family)

Members of the MIP family are mainly involved in facilitating the transport of both water and small neutral solutes through the cellular membrane in all domains of life. There are about six MIP subfamilies, the two major being the aquaporins (AQPs) and the glycerol-uptake facilitators (GLPs) (Zardoya and Villalba, 2001). The AQP and GLP subfamilies were used for the initial validation of the SDP-pred method (Kalinina et al., 2004). An arbitrary measure for functional significance of a site used by Kalinina et al. (2004) was the proximity to the glycerol molecules that are bound inside the GLP pore channel in the crystal structure 1FX8 (Fu et al., 2000). Sites that were conserved in the training set of sequences were excluded. This scenario yielded a set of 23 putative functional sites closer than 5 Å to any of the three glycerol molecules (Kalinina et al., 2004).

5.3.4 Sequence retrieval and alignment

R-Smad protein sequences were collected using the NCBI query for sequence retrieval (`www.ncbi.nih.gov`). This resulted in 15 non-redundant sequences for AR-Smads and 17 for BR-Smads. All sequences were aligned using the PSIPraline multiple sequence alignment online server (`www.ibi.vu.nl/programs/pralinewww`) (Simossis et al., 2005; Simossis and Heringa, 2003). From the alignment obtained, the MH2 domain was selected for further analysis.

For the Ras superfamily, as used by del Sol Mesa et al. (2003), sequences and alignments were directly obtained from Pfam-B (Finn et al., 2006). Selection of sequences for Ras, Ral, Rab5 and Rab6 families was simply performed by matching sequence names on Ras, Ral, Rab5 and Rab6, respectively yielding 69, 20, 4 and 6 sequences. The hypervariable termini of Rab5 and Rab6 were not present in the Pfam alignment.

For the MIP family, we took all bacterial AQP and GLP protein sequences, as defined in Figure 1 of Kalinina et al. (2004). For the other sequences mentioned in this figure, the classification was less obvious. We therefore decided not to take these into account. This scenario yielded 12 sequences for the AQP subfamily and 48 for both GLP subfamilies (one GLP identifier 'YA17_HAEIN' could not be resolved). The sequences were aligned using PSI-Praline as mentioned above for the Smads.

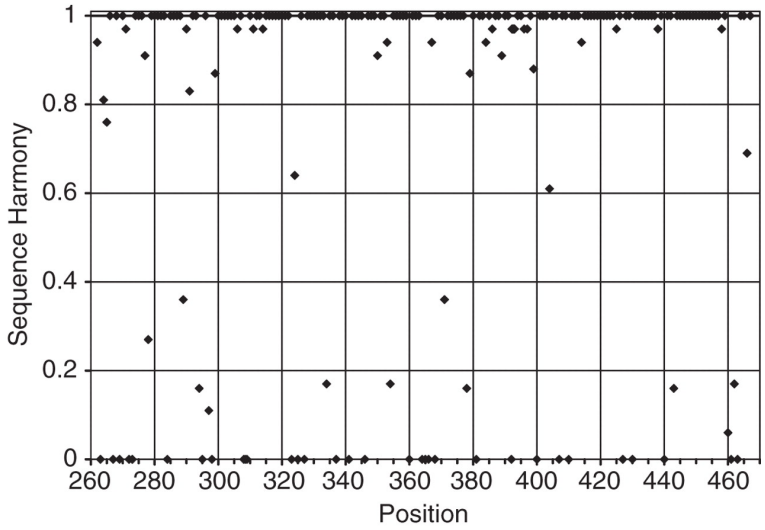


Figure 5.1: SH for AR-Smads versus BR-Smads along the sequence of the MH2 domain of Smads. 40 sites have SH zero. Most sites have SH one, meaning these have the same composition in each subgroup. Relatively few sites have intermediate SH values.

5.4 Results

5.4.1 TGF-β-associated transcription factors (Smad family)

SH between AR- and BR-Smads was calculated using Equation 5.3 for all 211 positions along the Smad MH2 alignment, as shown in Figure 5.1. It is clear that the vast majority is conserved overall, *i.e.* 135 sites have SH = 1. On the other hand, relatively few sites are completely non-harmonious, *i.e.* 32 have SH = 0. Out of these, only 13 are conserved in both groups, as can been seen in Table 5.2. The other 19 sites show a more variable composition but the subgroups still have non-overlapping compositions. Further, a relatively small fraction of sites show intermediate SH values, *i.e.* 44 are in between 0 and 1. In the range of 00.2, eight sites are found, and between 0.2 and 0.8 only seven. This means that the choice of a cut-off value for determining sites of 'low' SH is not very critical in this case.

Out of the 211 residues in the Smad MH2 sequence alignment, only 40 have a low SH (SH $\leq$ 0.2). In Table 5.3, these sites are listed together with their known interactions. It is clear that the vast majority of low-harmony sites also have a known function. sumMH2spec provides a further summary of the number of low-harmony sites associated with a particular function. Out of the 40 low-harmony sites, 27 have a known function (68%), while of the 32 non-harmonious sites (SH = 0) 23 have a known function (72%). In total, there are 29 known sites of functional specificity in

Function	SH = 0	SH ≤ 0.2
TβR-I/BMPR-I/ALK1/2 binding	8	10
c-Ski/SnoN binding	5	7
SARA/Mixer/FAST1	10	10
Total 'functional'	23	27
Putative function	8	10
Unknown function	1	3
Total 'unknown'	9	13
Total	32	40
Functional versus total (%)	72	68
Functional + putative versus total (%)	97	93

Table 5.3: Summary of functional sites and sites of unknown function with no (SH zero) or low (SH < 0.2) SH and specificity of the functional prediction.

the Smad dataset. These include several with rather high compositional variation in one or both groups (Table 5.2). Of the 171 remaining sites (SH > 0.2), only two sites are known to be important for the specificity of the Smadreceptor interactions (1%).

The AMAS method selects six sites that are different with respect to one or more physiochemical properties (Table 5.2). Consequently, these sites have non-overlapping amino acid compositions between the groups. Three are conserved in both groups, two are conserved in one group and one is not conserved in either group. Five out of the six selected sites are of known function, but clearly the majority of the 29 known functional sites are missed.

TreeDet/MB selects 21 sites that are completely conserved in at least one group but show no intergroup overlap (Table 5.2). Seven additional sites show the same characteristics but are not selected. The 21 selected sites contain 16 known functional sites. The other 7 sites contain 5 known functions.

SDP-pred selects 12 sites that are conserved in both groups and show no overlap (Table 5.2). Of these, 9 have known functions. One site of known function is conserved in both groups, but is not selected by the method. SDP-pred reaches a maximum coverage of ∼80%. All sites selected by SDP-pred are also selected by TreeDet/MB (see above).

All sites selected by AMAS, TreeDet/MB or SDP-pred are also selected by SH. SH selects 18 additional sites, 11 of which have a known function. None of the 18 sites are conserved in both groups, which explains why the other methods have difficulties finding them. Two further known functional sites remain undetected by any of the included methods.

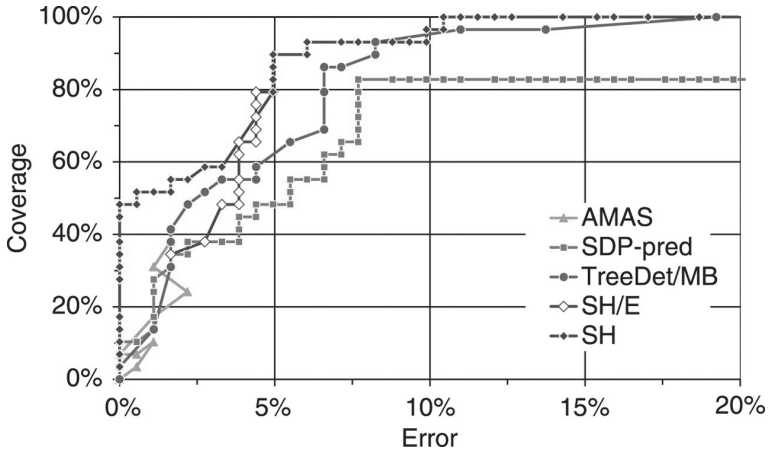


Figure 5.2: ROC plots for AR-Smads versus BR-Smads using the different prediction methods. For SH and SDP-pred the ranked results were used. One point for each unique rank value is drawn. Coverage is calculated as TP/(TP + FN), and Error as FP/(TN + FP). Note that the error rate is shown up to 20%. Note also that no method reaches a higher coverage for higher error rates.

The ROC plot shows that TreeDet/MB, SDP-pred and SH/E all reach ∼40% coverage with similar error (Figure 5.2). At higher coverage, TreeDet and SH/E yield lower error than SDP-pred. AMAS does not reach higher coverage and therefore it has not been applied to the other test-sets. It is clear that SH outperforms its counterparts at all coverage/error combinations. Notably, the first 14 sites selected by SH are all validated functional sites.

5.4.2 Small GTPases (Ras superfamily)

In Figure 5.3, ROC-curves are shown for two sets of families from the Ras superfamily of small GTPases; Rab 5 versus Rab 6 (Stenmark and Olkkonen, 2001; Stenmark et al., 1994), and Ras versus Ral (del Sol Mesa et al., 2003; Bauer et al., 1999).

For Rab5/6 specificity (Figure 5.3A), the SH predictions show a very high coverage, even at low error rates. Overall, TreeDet/MB, SDP-pred and SH/E achieve somewhat similar coverage and error. Nevertheless, SH/E reaches 35% coverage at only half the error of SDP-pred. At that coverage, TreeDet/MB shows a 2-fold error relative to SDP-pred. SH overall outperforms all other methods by a significant margin. The first 20 sites selected contain only two unknowns, while 70% coverage is attained within 10% error (Figure 5.3A).

For the Ras/Ral test-set, however, the prediction methods perform more similarly (Figure 5.3B). The first four sites (33%) selected by SDP-pred are validated. At higher coverage, SH/E performs slightly better than SDP-pred. Nevertheless, SDP-pred reaches 100% coverage at ∼12% error, while at this error rate SH/E and SH are

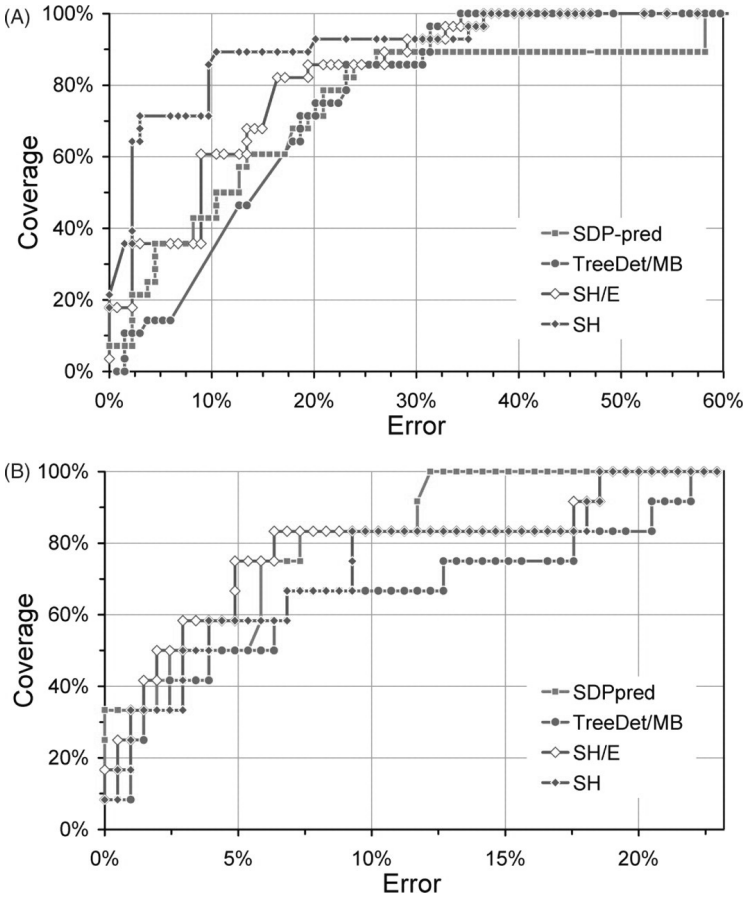


Figure 5.3: ROC plots for (**A**) Rab 5/6, and (**B**) Ras/Ral-specific sites using SH, SDPpred and TreeDet/MB (see text and Figure 5.2 caption for details). Validation of Rab5/6-specific sites was taken from Stenmark and co-workers (Stenmark and Olkkonen, 2001; Stenmark et al., 1994), and for Ras/Ral specificity from Bauer et al. (1999) and (6,20) and del Sol Mesa et al. (2003). Note that error rate is shown up to 60% and 23% for (A and B), respectively.

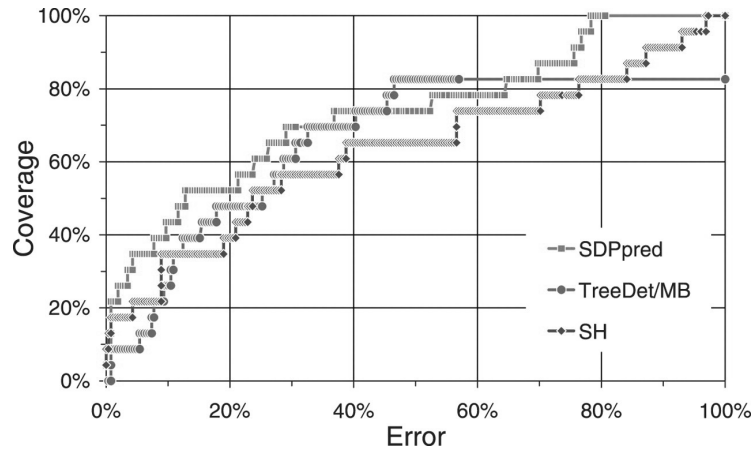


Figure 5.4: ROC plots for MIP specificity using SH, TreeDet/MB and SDPpred (see text and Figure 5.2 caption for details). MIP subfamily-specific sites were selected based on a minimum distance of 5 Å from the glycerol molecules bound in the pore channel in the glycerol uptake protein crystal structure 1FX8 (Fu et al., 2000).

just above 80% coverage and TreeDet/MB attains 65% coverage.

5.4.3 Integral membrane transporters (MIP family)

Figure 5.4 shows the relative performance over the MIP family. Predictions with the SDP-pred server were obtained using our own alignment and are comparable with those reported by Kalinina et al. (2004).

At 20% error, all methods only achieve medium coverage (∼50%, Figure 5.4). In contrast, for the other test sets at 20% error all other methods achieve higher coverage (between 75% and 100%). SH achieves somewhat higher coverage at the lowest error rates (the first 3 selected sites are <5 Å from their nearest ligand). At higher error rates, SDP-pred generally outperforms the other methods. TreeDet/MB performs similar to SH at lower coverage and more similar to SDP-pred at higher coverage. For this dataset, ranking for SH was dominated by the harmony score and only minute differences were seen between SH and SH/E (data for SH/E not shown).

5.4.4 Spatial and functional clustering

In Figure 5.5, the SH data are projected onto representative crystal structures of the four protein families included in the benchmark.

For the Smad2 MH2 domain, we identified a limited number of spatial clusters of low-harmony sites as indicated in Figure 5.5A. Taking membership of these clusters as a guideline, we assign putative functions to 10 out of the 13 unknowns, as indicated in Table 5.2. It is clear that most of these could not have been assigned from the sequence alone. Unknowns 392, 400, 407 and 410 can be assigned a putative function in FAST1 and/or Mixer binding. The two SARA-binding residues (366 and 368) in this FAST1/Mixer/SARA-binding cluster are furthest away from the unknowns. Importantly, 407 is known not to be involved in receptor interactions (Lo et al., 1998). Unknowns 334 and 337 can be assigned a putative function for Co-repressor (c-Ski/SnoN) binding. Position 337 is known not to be involved in SARA binding (Wu et al., 2000). Unknowns 269, 272, 273 and 443 can be assigned a putative function in SARA binding. In all, only three sites remain that cannot be assigned a putative function, out of a total of 40 low-harmony sites.

For the other test-sets, we have chosen layouts in Figure 5.5 analogous to those used in the corresponding papers (see Figure 5.5 caption). For Rab5/6 (Figure 5.5B) and Ras/Ral (Figure 5.5C), it can be seen that selected sites form localized groups in the structure. For the MIPs (Figure 5.5D), this trend is much less salient.

5.5 Discussion

In this paper, we have introduced the SH as a means to pinpoint putative functionally different sites and compared the results with other methods that rely mostly on conservation.

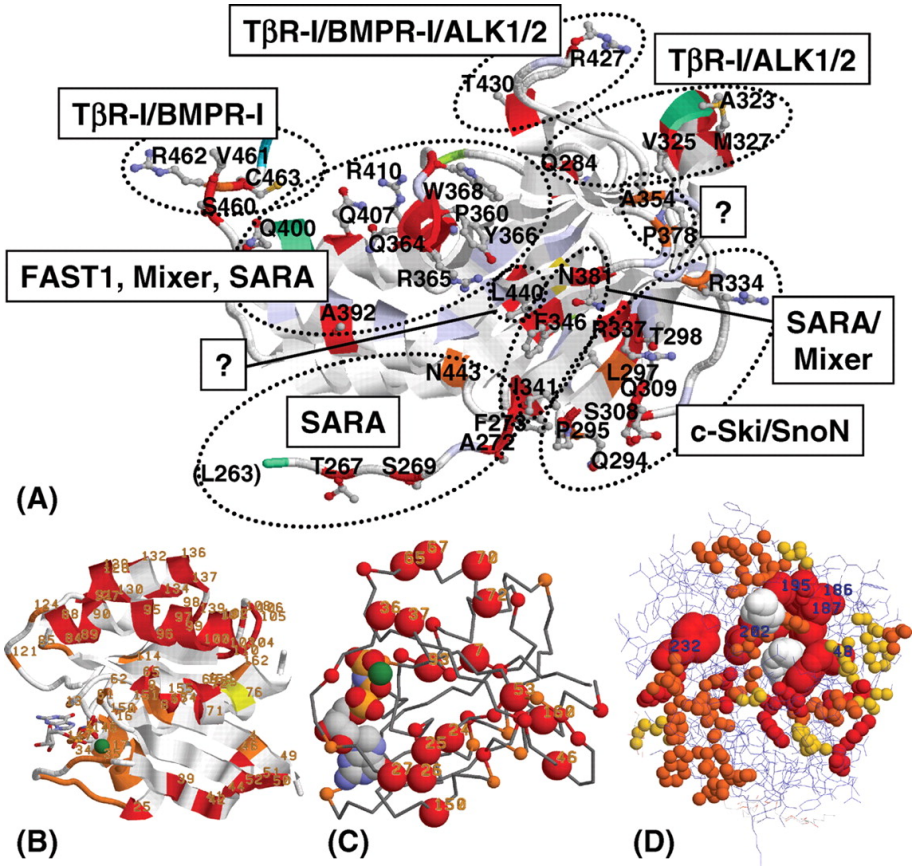


Figure 5.5: SH in a representative crystal structure for each of the test-sets. Non-harmonious sites (SH zero) are red and low harmony (SH $\leq$ 0.2) orange. Residue numbers for the low-harmony sites (SH $\leq$ 0.2) are indicated. (**A**) AR-Smads versus BR-Smads colour-coded onto the crystal structure of the MH2 domain of Smad2 (1KHX) (Wu et al., 2000). The spatial clustering of low-harmony sites is indicated with dotted ellipses, and clusters are labeled with corresponding known functions. Intermediate values go from white to light blue for maximum harmony (SH one). (**B**) SH for Rab5/6 using the crystal structure 5P21 and a representation and orientation similar to Figure 3a in Stenmark and co-workers (Stenmark and Olkkonen, 2001; Stenmark et al., 1994). (**C**) *id.* for Ras/Ral using 5P21 and similar to Figure 4 in del Sol Mesa et al. (2003). (**D**) *id.* for MIP using 1KHX and similar to Figure 5 in Kalinina et al. (2004).

5.5.1 Performance varies for methods and datasets

For the Smad test-set, all five methods reach >80% coverage at <10% error (Figure 5.2). Nonetheless, differences among the methods are substantial and SH maintains highest coverage virtually throughout. For the Rab5/6 test-set, all four methods perform at least moderately well (Figure 5.3A). The differences among the methods are somewhat larger than for the Smad test-set, with coverage ranging from 40% to ∼80% at 10% error. Also here, SH performs best.

As to the Ras/Ral test-set, all four methods perform well and reach >80% coverage at 10% error (Figure 5.3B). The differences among the predictions are small and there is no method that outperforms the others over a pronounced range in coverage or error. For the MIP test-set, overall prediction quality attained by the four methods is dramatically lower than for the other test-sets, with coverages ranging from 20% to 40% at 10% error (Figure 5.4). Although the differences are relatively small, SH performs slightly worse at a coverage >35% than its counterparts for this dataset.

In summary, for two of the four test-sets all predictions are very accurate while for another set only SH achieves high accuracy. In contrast, for the fourth set none of the methods give accurate predictions. Furthermore, while predictions for two of the four test-sets show only small performance differences, the differences are larger for the remaining two test-sets and here SH achieves highest coverage throughout.

5.5.2 Different methods select different sites

AMAS focuses on conservation of physiochemical properties. This principle leads to very specific predictions but the existing method seems to be overly conservative.

SDP-pred, TreeDet/MB and SH/E focus on conservation and yield good predictions in general. Differences among these methods are relatively small. This suggests that the signal arising from conservation dominates over possible differences arising from methodology or ranking schemes. However, there appears to be a margin for improvement that could indicate that other factors are more important than conservation for determining functional specificity.

SDP-pred uses the Bernoulli estimator to automatically determine an optimal cut-off and yields highly specific but conservative predictions. TreeDet uses an internal algorithm for unsupervised grouping of sequences that may not always lead to finding differences of interest.

SH focuses on non-overlapping composition between subfamilies and yields good predictions with very high coverage in several cases. This suggests that subfamily differences and sequence context are crucial determinants for functional specificity.

The examples studied here indicate that emphasis on conservation is not sufficient to specifically detect known functional sites. Shifting the focus completely to differences as we have implemented in the SH method, seems to give better predictions overall, at least on the datasets tested here. However, it remains to be seen whether other factors may be involved and what relative weight should be attached to conservation and compositional differences.

The difference between our SH and SH/E methods is the use of sequence context by SH for the ranking of selected sites. Generally, SH performs better than SH/E, which indicates that sequence context may be an important indicator to select regions of interest. It is interesting to note that several of the sequential ranges identified in the Smads (Table 5.2) are located inside helices. This may seem counterintuitive since first and second neighbours are on opposite sides in a helix. Nevertheless, the average distance between the neighbouring C_βs in a helix is only $\sim$5 Å and the flexibility of the sidechain would allow them to approach close enough to participate in the same function. The alternating pattern of β-strands is accommodated by the inclusion of second neighbours, as already described. An interesting refinement of the method could integrate knowledge of the secondary structure, *e.g.* selecting only first neighbours for loops, only second neighbours for b-strands, and first, third or fourth neighbours for helices.

5.5.3 Performance varies for different datasets

For the protein families presented here, the functional specificities of many sites have been investigated. Validation based on point mutations (Smad, Ras/Ral) is the most specific, but ignores the possible cooperative role of additional sites. Validation based on the exchange of sequence segments (Rab5/6) does include the possible cooperative role of sites, but makes it difficult to assess the discriminatory effect of individual sites. Validation based on distance-to-ligand (MIP) provides no information about the individual role of sites. In all test-cases included, many sites remain that have not been investigated experimentally and it is likely that many additional functional sites exist. This would lead to an underestimation of the number of true positives and complicates the evaluation of different prediction results.

For the two test-sets based on point-mutants (Smad, Ras/Ral), we see a generally high performance of the methods. The abundance of direct and high quality experimental validation for the Smad test-set allows an accurate assessment of the quality of the predictions. For the other two test-sets (Rab5/6, MIP), performance is generally lower. The methods perform well in identifying specific sites, but experience more difficulty in delineating regions corresponding to swapped segments or residues that are close to bound ligands.

The degree of conservation differs significantly between sites, but the available experimental evidence does not necessarily cover this distribution uniformly. Notably, conserved sites are often likely to be functionally relevant and this has led to a likely bias in studying mutations of relatively conserved sites. Such a bias would lead to an overestimation of the importance of conservation.

For further development of this field of research, a well crafted collection of high-quality experimental data for a variety of protein families would be of great value. Preferably, experimental validation would be based on a representative mix of sites with different degrees of conservation and include many specific point mutations as well as swapped segments to assess supporting roles of sites and strengthen confidence in true negatives.

5.5.4 Functional sites can be grouped by spatial clustering

Functionally related residues tend to form spatial clusters in a protein structure. SH selects sites of unknown function for the protein families considered, and most of these cluster with sites of known function. This provides a way of grouping selected sites and can be used to transfer functional annotation for residues assigned to the same cluster. Preliminary data seem to indicate that coarse-grained structures of medium or even low quality, *e.g.* by homology modeling or *ab initio* prediction, may also prove sufficient for this type of clustering.

For the Smad protein family many important questions about the specific interactions with other factors in the TGF-β and BMP-associated pathways are still open. The sites selected by SH are likely candidates for supporting these specific interactions. The putative functions assigned to these sites based on the spatial clustering may provide important guidance for future experiments.

5.6 Conclusion

We have shown that SH achieves predictions of high quality. While some other methods use sophisticated statistical analysis methods, this does not appear to lead to an increased quality of the predictions. The simplicity of SH makes the results easy to understand. SH achieves high coverage in general and selects additional sites of unknown function. The location of these sites in the crystal structures associated with the benchmark sets used here indicates most of them as promising candidates for further investigation.

The current study provides, to the best of our knowledge, a first attempt of a systematic comparison of prediction methods for functional differences between protein subfamilies. From this analysis, we conclude that exploiting conservation alone is not sufficient, and that more emphasis on sequence differences and context could be the crucial factor for identification of sites of functional specificity. The SH web-server is available at `www.ibi.vu.nl/programs/seqharmwww`.

5.7 Acknowledgements

K.A.F. and W.P. acknowledge financial support from the Netherlands Bioinformatics Centre, BioRange Bioinformatics research programme (SP 2.3.1 and SP 3.2.2, respectively). We thank the two anonymous referees for their thoughtful and thorough comments that have lead to significant improvements in the paper. We thank Willie Taylor for a critical reading of the manuscript. We also thank Mikhail Gelfand for sharing alignment data on the MIP family. Funding to pay the Open Access publication charges for this article were waived by Oxford University Press.

CHAPTER 6

Sequence harmony: detecting functional specificity from alignments

Published as:

Feenstra, K.A., Pirovano, W., Krab, K., and Heringa, J. (2007).
Sequence harmony: detecting functional specificity from alignments.
Nucleic Acids Res, 35(Web Server issue):W495–W498.

Abstract

Multiple sequence alignments are often used for the identification of key specificity-determining residues within protein families. We present a web server implementation of the Sequence Harmony (SH) method previously introduced *[Nucl Acids Res 34 6540]*. SH accurately detects subfamily specific positions from a multiple alignment by scoring compositional differences between subfamilies, without imposing conservation. The Sequence Harmony web server allows a quick selection of subtype specific sites from a multiple alignment given a subfamily grouping. In addition, it allows the predicted sites to be directly mapped onto a protein structure and displayed. We demonstrate the use of the SH server using the family of plant mitochondrial alternative oxidases (AOX). In addition, we illustrate the usefulness of combining sequence and structural information by showing that the predicted sites are clustered into a few distinct regions in an AOX homology model. The Sequence Harmony web server can be accessed at `www.ibi.vu.nl/programs/seqharmwww`.

6.1 Introduction

During the past years there has been a wide interest in studies of specificity-determining residues within protein subfamilies (Whisstock and Lesk, 2003). Consequently, an increasing number of methods and web applications has become available that offer functional analyses of subtype specificity from multiple alignments (Mirny and Gelfand, 2002; Kalinina et al., 2004; Donald and Shakhnovich, 2005; Ye et al., 2006).

Previously we evaluated advantages and limitations of several of the state-of-the-art methods and introduced a new method called Sequence Harmony (SH) (Pirovano et al., 2006). SH accurately detects positions within an alignment that are responsible for functional differences between two protein subfamilies. To further facilitate the use of SH for a broad audience, we have implemented a comprehensive web server. The SH web server offers a fast, one-step analysis with a number of options, yielding results that can be interpreted easily.

In this paper we will guide the user through all the steps of the SH web-application by means of a biologically relevant example. We will look for subtype specific sites for the two subfamilies of the AOX protein family of plant alternative oxidases. The subtype specific sites found are the best candidates to explain the functional differences. Other relevant applications of this method include pathway specificity, ligand specificity, host-specific viral infection, viral disease progression differences and viral drug resistance.

6.2 Methods

6.2.1 Implementation

The Sequence Harmony method, as described previously (Pirovano et al., 2006), is currently implemented as an awk program. The main steps are as follows. Sequences are read from the alignment and separated into two user-specified groups. For each group separately and combined entropies (E) are calculated. SH values are calculated as $SH = \frac{1}{2}(E_{A+B} - E_A - E_B)$, with $E_{A+B} = -\sum(p_A + p_B)\log(p_A + p_B)$, *i.e.* using the sum of the normalized frequencies of groups A and B. SH values range from zero for completely non-overlapping residue compositions, to one for identical compositions. Next, sites are selected that have a SH score below a cutoff. Stretches of neighbouring selected sites are identified and the size of each of these stretches is assigned to the sites as the 'Rank'. Finally, selected sites are sorted on *i)* increasing SH, *ii)* decreasing Rank, and *iii)* increasing entropy. This sorted list of selected sites is the primary result of the SH algorithm. Currently alignments of up to about 5 million residues (including gaps) can be processed with runtimes on the order of tens of seconds, excluding the generation of the high-quality output PyMol image.

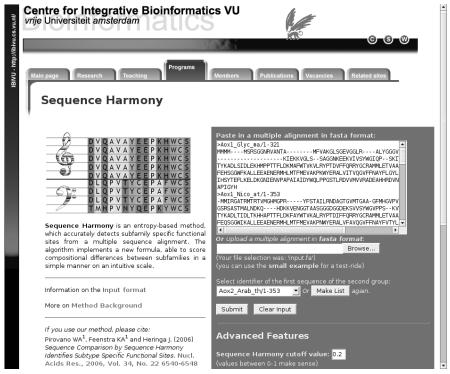
The separate chains of the optional protein structure (PDB file) are aligned with the input alignment using the 'profile' option of Muscle (Edgar, 2004c). Subsequently, the chains with the highest average Sequence Harmony score with respect to the input alignment are selected for graphical display as an image in an interactive Jmol applet.

The SH web server contains basic and advanced options. Sanity of the input options is checked, *e.g.*, whether an input alignment is uploaded or pasted, and not both, whether it is proper FASTA format. Problems are reported with an error message and offending input fields are highlighted. An example screenshot of the main page and input form is shown in Figure 6.1. The main input is a multiple sequence alignment of the protein family and a subdivision into two groups. Sequence data in FASTA format is case-insensitive and may be split over multiple lines; gaps should be represented with a dash ('–'); and lines beginning with a semi-colon (';') are ignored as comment. From the alignment the program automatically generates a list of sequence IDs. The user defines the two input groups by selecting the identifier of the first sequence of the second group. By default, a cutoff of 0.2 for the SH score is used.

6.2.2 Use of the Server

Advanced features allow more control over the analysis and output, but the default settings usually suffice for a basic analysis. The SH cutoff can be adjusted between zero (allowing no compositional overlap) and one (allowing full overlap). A higher cutoff will lead to the selection of a larger number of sites. A reference sequence can be selected from the alignment and a starting position can be set to provide a reference numbering in the output tables. Additionally, a PDB identifier can be specified by its four-letter code, to visualize the selected sites in the protein structure. Alternatively,

(A)



(B)

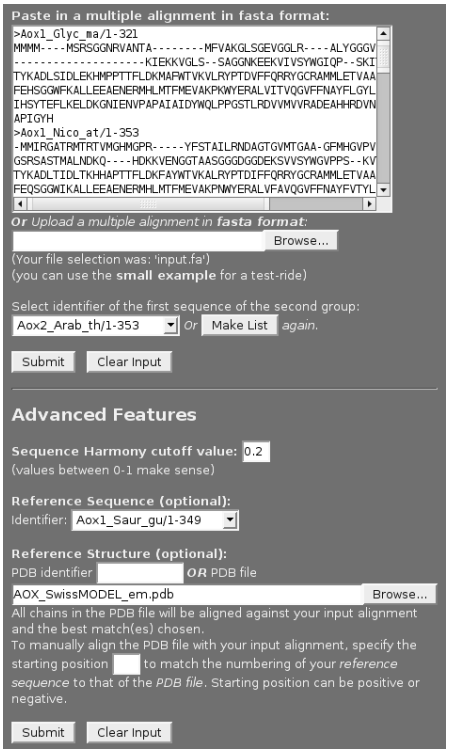


Figure 6.1: Input screen of the Sequence Harmony server, showing part of the AOX alignment and model input. **(A)** Overview of the Sequence Harmony web page; and **(B)** the Sequence Harmony input options and advanced features.

a PDB file can be uploaded. The PDB file is automatically aligned with the multiple alignment, or manually when the reference position is set.

The output is a sorted table and optional graphical representations of the selected sites as an image (generated by PyMol, `www.pymol.org`) and as an interactive Jmol applet (Herraez, 2006). The Jmol applet includes buttons to set the SH cutoff for displaying selected sites, and to show or hide non-aligned chains in the PDB file. An example screenshot of the main output page is shown in Figure 6.2. An additional, non-sorted, table is available that lists all sites in the alignment. The tables are coloured from red (predicted subtype specific sites) to blue (sites not predicted).

6.3 Results

6.3.1 The AOX family of alternative plant oxidases

The cyanide-insensitive plant alternative oxidase (AOX) is a mitochondrial non-heme quinol oxidase (Affourtit et al., 2002). An established function of AOX is thermoge-

nesis in the spadices of *Aroid* lilies, but little is known about AOX function in other tissues (Wagner and Moore, 1997). AOX is encoded in two discrete gene subfamilies (Considine et al., 2002). AOX1 is found in monocot and dicot plants and is induced by stress stimuli (Considine et al., 2002; Vanlerberghe and McIntosh, 1997). AOX2 is usually constitutive and has at present not been found in monocot plants (Considine et al., 2002).

We retrieved 31 AOX sequences from the NCBI non-redundant database (`www.ncbi.nih.gov`); 24 of the AOX1 and 7 of the AOX2 subfamily. Sequences were aligned using PSI-Praline (Simossis and Heringa, 2003; Simossis et al., 2005) using secondary structure prediction information from PSIPRED (Jones, 1999). The alignment is available as supplementary material. Lacking an experimental AOX structure, we constructed a homology model using of the catalytic iron-binding N-terminal domain. Swiss-Model (Guex and Peitsch, 1997) was used to align the *Sauromatum guttatum* AOX1 sequence with the template structure 1AFR (Lindqvist et al., 1996), following the description of Andersson and Nordlund (1999). The PDB file is provided as supplementary material.

6.3.2 Sample input

The multiple alignment obtained was uploaded in the appropriate box of the SH web server and the first AOX2 sequence (AOX2_Arab_th) was selected as the first sequence of the second group, see Figure 6.1. The default SH cutoff of 0.2 was used, meaning that partial overlap in residue composition is allowed. We chose AOX1 from *Sauromatum guttatum* as reference sequence so that the subtype specific sites obtained will be numbered accordingly (Andersson and Nordlund, 1999). The PDB file of the model structure was uploaded to the SH web server.

6.3.3 Sample output

The results page of the Sequence Harmony server (see Figure 6.2) shows at the top a summary of the input parameters, and optionally the graphical representations of the protein structure. Below that is the table containing all sites with a SH score (SH) below the cutoff value (default 0.2). These are sorted first on increasing SH score, then on decreasing rank (Rnk), number of neighbouring sites below the cutoff) and finally on increasing total Entropy (`AB`) of the alignment position, as described previously (Pirovano et al., 2006). Position Ali and Ref show the sequence position in the alignment and reference sequence, respectively. The 'Consensus' columns give all residue types present in groups A and B, respectively, in order of decreasing frequency and in lowercase when the frequency is less than half of the highest. In addition, a link is provided to the raw table, that holds all information of the complete alignment without selection or ranking. The raw table is available as supplementary material.

For the AOX family we found 9 sites with a SH value of zero (bright red in Figure 6.2). An additional 8 sites have SH values between zero and 0.2 (from light red to white). In the 'Rnk' column, a value of 2 signifies a stretch of two consecutive

Sequence Harmony for 'AOX_PSIPral-AliSorted.fasta'

Generated on Tue 27 Mar 2007 15:08 by SeqHarm version 1.1
from the Sequence Harmony Webserver at http://www.ibi.vu.nl/programs/seqharmwww/
Please cite:
Walter Pirovano*, K. Anton Feenstra* and Jaap Heringa
"Sequence Comparison by Sequence Harmony Identifies Subtype Specific Functional Sites"
Nucl. Acids Res., 2006, Vol. 34, No. 22 6540-6548
*joint first authors.
Found reference sequence 'Aox1_Saur_gu/1-349', offset=1.
Group A: 24 sequences, length 366
Sequences: Aox1_Glyc_ma/1-321, Aox1_Nico_at/1-353, Aox1_Saur_gu/1-349, Aox1_Sola_tu/1-34
Group B: 7 sequences, length 366
Sequences: Aox2_Arab_th/1-353, Aox2_Cucu_sa/1-346, Aox2_Mang_in/1-274, Aox2a_Glyc_ma/1-
Cutoff: 0.2.

Archive of all output files

Results for AOX_PSIPral-AliSorted.fasta

Aligned chains from AOX_SwissMODEL_em.pdb or together with whole alignment

Selected 17 positions below cutoff (0.2)

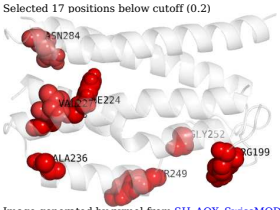
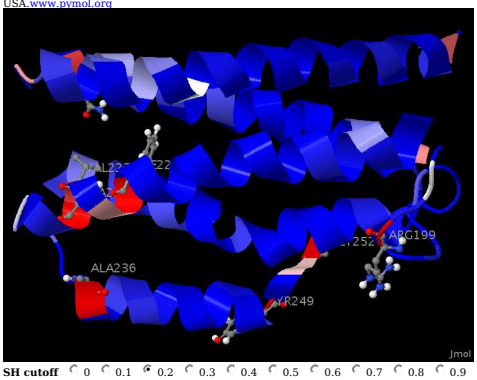


Image generated by pymol from SH_AOX_SwissMODEL_em.pdb pdb file with '1-SH' as B-factors and using display.pml pymol script:
DeLano, W.L. The PyMOL Molecular Graphics System (2002) DeLano Scientific, Palo Alto, CA, USA.www.pymol.org



SH cutoff 0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Show Whole structure Only Aligned

Raw Table

Selected 17 positions below cutoff (0.2)

SH: 0.00 0.04 0.08 0.12 0.16 0.20 0.36 0.52 0.68 0.84 1.00

Position		Entropy				SH	Rnk	Consensus	
Ali	Ref	A	B	AB	rel.			A	B
245	A224	1.06	0.00	1.59	1.26	0.00	2	Astm	V
216	R199	0.00	0.00	0.77	1.26	0.00	1	R	K
241	F224	0.00	0.00	0.77	1.26	0.00	1	F	M
141	R124	0.00	0.00	0.77	1.26	0.00	1	R	M
253	A236	0.00	0.00	0.77	1.26	0.00	1	A	L
266	Y249	0.00	0.00	0.77	1.26	0.00	1	Y	F
136	R119	0.74	0.00	1.34	1.26	0.00	1	Kr	P
269	G252	1.06	1.38	1.90	1.26	0.00	1	Gat	Lfc
354	L337	2.05	0.00	2.36	1.26	0.00	1	Mqrhl	K
122	P107	1.36	0.59	1.89	1.17	0.07	2	Pqat-	Rt
244	V227	0.25	0.00	0.82	1.10	0.13	2	Vl	L
170	R153	0.90	0.00	1.33	1.10	0.13	1	Wlry	R
301	N284	1.67	1.15	2.20	1.07	0.15	1	AKnd	Sen
55	Q48	3.13	2.52	3.64	1.07	0.15	1	V-armlpqstw	Fcghy-
147	Q130	2.03	1.45	2.50	1.04	0.17	1	KTesqr	HRq
101	V86	2.73	1.38	3.02	1.03	0.18	1	AESvg-kt	Knt
121	P106	2.67	1.66	3.07	1.02	0.20	2	EPQdkav-	Sety

Figure 6.2: Output screen of the Sequence Harmony server with results of the AOX analysis. Included are a summary of input parameters, a PyMol rendered image, the Jmol applet and controls, and the output table. Links to the raw table with all alignment positions and to additional output files are also provided.

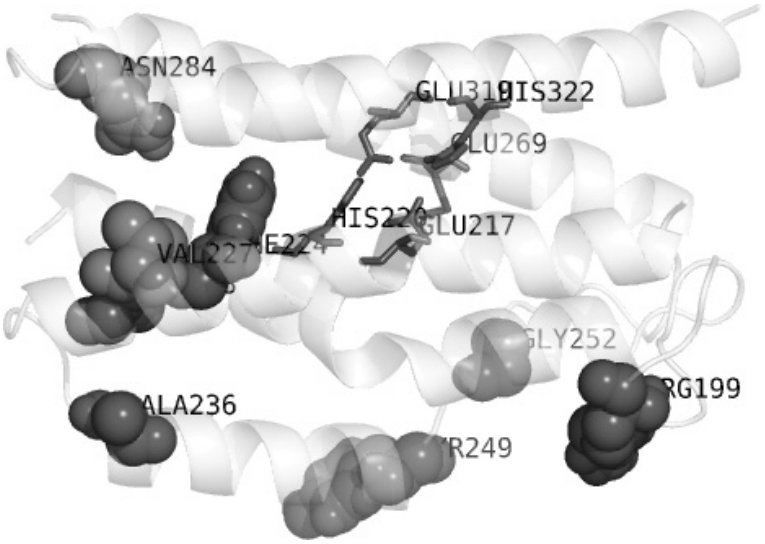


Figure 6.3: Model of the AOX iron-binding domain with low-harmony sites shown in spheres and the iron-binding Glu and His residues in sticks. Figure created using PyMol (`www.pymol.org`).

sites with SH values below the cutoff (alignment positions 244–245 and 121–122). Interestingly, out of nine sites with zero harmony, four have non-zero entropy in at least one group, *i.e.* they are not totally conserved within both groups (alignment positions 136, 245, 269, 354). This is also reflected in the 'Consensus' columns, which show multiple residue types occurring.

The graphical representations of the protein structure show the predicted sites in our homology model of the C-terminal domain (positions 182–352), see Figure 6.3 and also Figure 6.2. This domain contains two sites of zero harmony and six of low harmony. Four low-harmony sites on the upper-left (Phe224, Val227, Ala228 and Asn284) are close to the Glu and His iron binding residues (Andersson and Nordlund, 1999) (Figure 6.3), and could conceivably play a role in ligand binding or activation mechanisms. Three other low-harmony sites, Arg199, Tyr249 and Gly252 are located around a putative membrane-binding region (Andersson and Nordlund, 1999), and could conceivably play a role in modulating the protein-membrane interactions or substrate access. The functional prediction of SH from the multiple alignment seems to be consistent with the structural prediction using homology modeling, although a different structural model could lead to different conclusions.

6.4 Conclusion

Our Sequence Harmony web server identifies putative subtype specific sites based on an alignment and selection of two groups as input only. In addition, sites can be easily linked to structural information using a structure from the PDB directly, or, as shown in our example, using a homology model of the protein. Using this structural information, selected subtype specific sites can be grouped into spatial clusters of sites that are likely to share functional relationships.

The combination of the Sequence Harmony algorithm and protein structural information has yielded a useful tool to interpret multiple sequence alignments, and to guide subsequent selection of interesting sites for experimental investigation.

6.5 Acknowledgements

K.A.F. and W.P. acknowledge financial support from the Netherlands Bioinformatics Centre, BioRange Bioinformatics research programme (SP 2.3.1 and SP 3.2.2) and the EC Network of Excellence 'ENFIN' (LSHG-CT-2005-518254). The Open Access publication charges for this article were waived by Oxford University Press.

6.6 Supplementary Material

The AOX input alignment, raw output table, 'coloured' PDB and PyMol script are available as supplementary material. In addition, the output is available at (`www.ibi.vu.nl/programs/seqharmwww/showcase_aox`).

CHAPTER 7

The meaning of alignment: lessons from structural diversity

Published as:

Pirovano, W., Feenstra, K.A., and Heringa, J. (2008).
The meaning of alignment: lessons from structural diversity.
BMC Bioinformatics, 9:556.

Abstract

Background: Protein structural alignment provides a fundamental basis for deriving principles of functional and evolutionary relationships. It is routinely used for structural classification and functional characterization of proteins and for the construction of sequence alignment benchmarks. However, the available techniques do not fully consider the implications of protein structural diversity and typically generate a single alignment between sequences.
Results: We have taken alternative protein crystal structures and generated simulation snapshots to explicitly investigate the impact of structural changes on the alignments. We show that structural diversity has a significant effect on structural alignment. Moreover, we observe alignment inconsistencies even for modest spatial divergence, implying that the biological interpretation of alignments is less straightforward than commonly assumed. A salient example is the GroES 'mobile loop' where sub-Ångstrom variations give rise to contradictory sequence alignments.
Conclusion: A comprehensive treatment of ambiguous alignment regions is crucial for further development of structural alignment applications and for the representation of alignments in general. For this purpose we have developed an on-line database containing our data and new ways of visualizing alignment inconsistencies, which can be found at `www.ibi.vu.nl/databases/stralivari`.

7.1 Introduction

Sequence comparison has become a major tool for biological research in the post-genomic era, forming the basis for functional annotation, classification, and analysis of evolutionary relationships. At the residue level, however, the relation between sequence, structure and function can often be obscure, and examples abound of proteins with a clear functional and homologous relationship but sharing negligible similarity at the sequence level.

Structural alignment therefore is the method of choice for reliable homology assessment and derived features like functional classification and phylogeny. This importance is reflected in the number of tools available for structural alignment, such as DALI (Holm and Park, 2000), SSAP (Taylor and Orengo, 1989), STRUCTAL (Gerstein and Levitt, 1998), MAMMOTH (Lupyan et al., 2005), CE (Shindyalov and Bourne, 1998) and COMPARER (Sali and Blundell, 1990) (for recent reviews on the topic, see Kolodny et al. (2005) and Mayr et al. (2007)). Databases for functional classification such as CATH (Orengo et al., 1997), FSSP (Holm et al., 1992) and PASS2 (Bhaduri et al., 2004) each derive directly from the use of one or more of these methods, whereas for SCOP expert input in the structural classification is deemed critical (Murzin et al., 1995). Structural alignments are also routinely used for benchmarking sequence alignment methods. A number of databases have been developed for this purpose, among which BAliBASE (Thompson et al., 1999), HOMSTRAD (Mizuguchi et al., 1998) and SABmark (Walle et al., 2005) are widely used. These

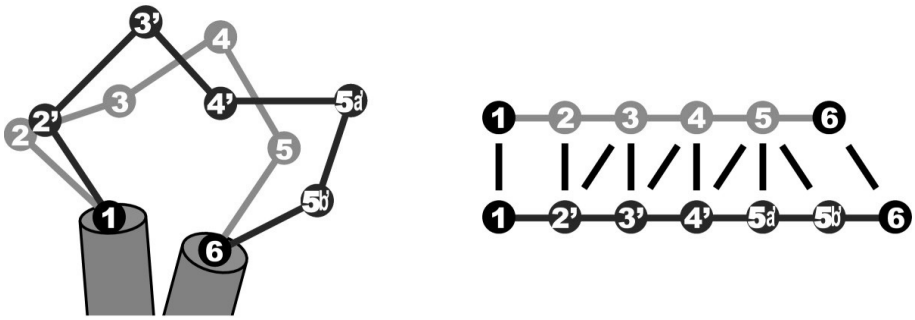


Figure 7.1: Dealing with structural flexibility: a single insertion (5', left) can lead to ambiguity in the pairwise residue alignment between the loops (right). Therefore, a simple one-to-one functional equivalence between residues from different proteins may not exist.

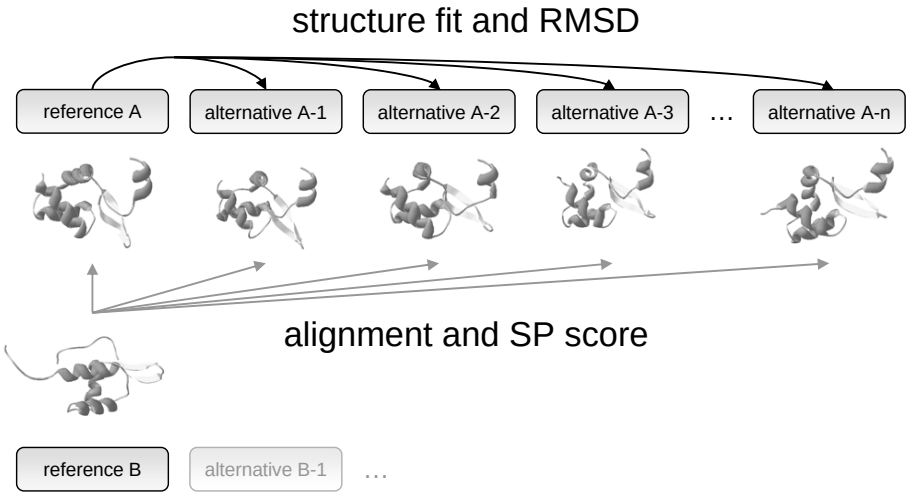


Figure 7.2: Overview of the approach. SP scores are calculated to describe the differences at the sequence level between the reference and alternative structural alignments. In addition each alternative structure (either obtained with molecular simulation or from the PDB) is fit onto the reference structure and root mean square deviations (RMSDs) are calculated.

databases often rely on expert knowledge and include a notion of 'core blocks', *i.e.* where alignment ambiguity does not occur and hence can be trusted. The general problem of uncertainty in sequence alignment has recently been discussed by Wong et al. (2008). Due to the complexity of interpreting non-trivial alignment regions, these are often omitted in large-scale evolutionary analyses, even though there is ample evidence for their fundamental importance (Wong et al., 2008; Rokas, 2008). An approach to pinpointing alignment ambiguity is the generation of ensembles of suboptimal alignments (Godzik, 1996), but computational demands remain prohibitive for genome wide studies.

Recent structural alignment methods have started to place emphasis on dealing with structural flexibility, such as FATCAT (Ye and Godzik, 2003), MultiProt (Shatsky et al., 2004), MATT (Menke et al., 2008) and RAPIDO (Mosca and Schneider, 2008). This may increase the consistency of alignments produced by each of these methods, but does not address the intrinsic ambiguity arising from structural divergence. The fundamental issue is whether a one-to-one equivalence exists between residues from different proteins that could be expressed as one definite alignment between sequences (Godzik, 1996). This is illustrated in Figure 7.1, where we show that a single insertion can lead to ambiguity in the functional correspondence between most residues in the loop.

To further elucidate the effect of structural diversity on structural alignment, we prepared two distinct comprehensive sets of alternative structures for proteins from the HOMSTRAD database of homologous protein families. The first set comprises proteins for which alternative crystal structures are available. The other set is derived from molecular dynamics simulations to explore a more extensive spectrum of possible structures. An overview of our analysis procedure is outlined in Figure 7.2.

Our main results show that in many cases structural variation strongly affects structural alignments, even for highly similar sequences. Moreover, the derived alignment appears to be highly sensitive to even small conformational changes of the proteins. The uncertainty in pairing up structural equivalent residues makes it difficult to determine which alignment alternative would describe most closely the functional relationship between the proteins. To address this issue, we show how alternative alignment visualizations may be used to exploit the information contained within variable alignment regions.

7.2 Results and discussion

7.2.1 Structural diversity and alignment stability

The relation between the variation of the alternative structures (RMSD) and the corresponding alignment similarity (SP score) is shown in Figure 7.3 (bottom panel). It is clear that the structural variation between crystal structures (in light grey) is much smaller (up to 3−4 Å RMSD) than that of the simulation snapshots (in dark grey; up to 10 Å RMSD). A crucial aspect is that even for small (<1 Å RMSD) and modest (13 Å RMSD) structural differences, alignments can easily vary up to

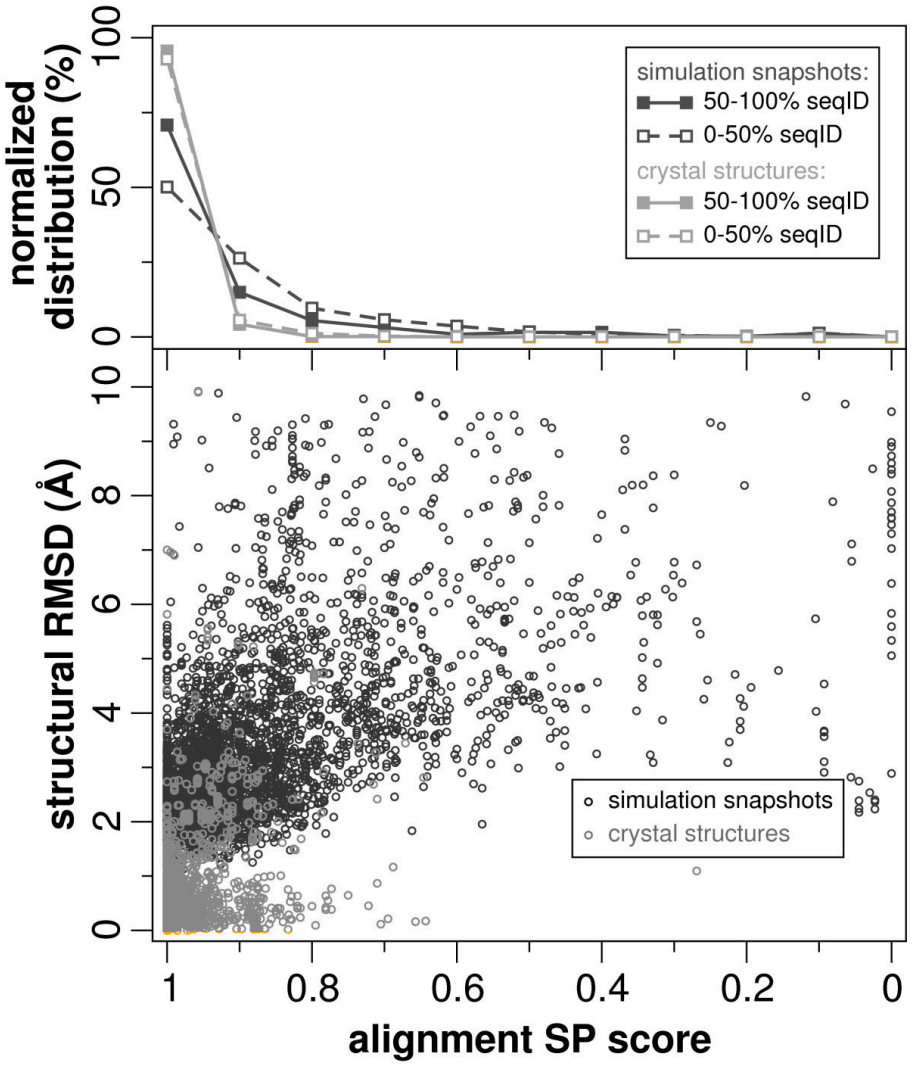


Figure 7.3: Effects of structure and sequence variation on the alignment. The bottom panel shows structural difference (measured by the RMSD) versus alignment similarity, measured by the SP score, which is defined as the ratio of the number of correctly aligned residue pairs in the test alignment to the number of aligned reference residues pairs that are reproduced in the query alignment. The top panel shows distributions of SP scores for alignments sharing less and more than 50% sequence identity. Light grey refers to alternative crystal structures while dark grey refers to alternative structures obtained from molecular simulations.

20% and sometimes as much as 40% or more in their SP score. On the other hand, a considerable number of alignments appear robust to larger (up to 6 Å RMSD) and even extreme (up to 10 Å RMSD) structural variations. Additionally, for the crystal structures, the sequence similarity has no effect on the variation in structural alignments (Figure 7.3, top panel). For the simulation snapshots, however, there seems to be a slight but distinct tendency for more similar sequences to have less variation in structural alignments, but this can be mainly attributed to the larger variations (>3 Å RMSD) in structure that arise from the simulations (see additional file 1, Section 7.6). As an alternative for RMSD measurements we also tested the rho-score (Maiorov and Crippen, 1995), a protein size-independent measure, which resulted in the same trend (data not shown).

A quite interesting example of the impact that small structural variations can have on the structural alignment is found in the GroES so-called 'mobile loop', which is the main region for interaction with GroEL and therefore is a crucial component of the GroEL/ES chaperonin machinery (Xu et al., 1997). The structural variations for this loop in *E. coli* GroES (Figure 7.4C, shown in blue) are almost negligible (whole protein C_α RMSDs 0.42 ± 0.13 Å). It is therefore surprising that the corresponding DALI sequence alignments with *M. tuberculosis* GroES show remarkable variation in this region (Figure 7.4A). To pinpoint the source of this variation, we also used three other structural alignment programs: CE, MATT and FATCAT. The latter two explicitly take structural flexibility into account and this leads to more consistent alignments in the variable loop (alignment positions 2069, Figure 7.4A). On the other hand, two regions (8489 and 107109, Figure 7.4A) are aligned consistently by DALI but show inconsistencies when aligned by CE and the two flexibility-aware methods. Strikingly, there is no overall consistency between the four methods, which is in line with several other studies where several structural alignment methods are compared (Kolodny et al., 2005; Mayr et al., 2007; Godzik, 1996). It should be stressed that the focus of this paper is not on comparing the performance of the various methods but rather on the effects of structural diversity. A comprehensive overview of the GroES variability is given by the alignment matrix and the consistency plots (Figure 7.4B). The alignment matrix scores the occurrence of aligned residue pairs over all alignments, similar to the dot-plot (Maizel and Lenk, 1981; Zuker, 1991). Consistency plots show for each residue the standard deviation from the alignment position of the consensus pair. The alignment matrix and associated consistency plots allow a detailed visualization of the variability while enabling easy interpretation of the ensemble of alternative alignments.

Although alignment uncertainty has been shown to have a great impact on large scale sequence analysis (Wong et al., 2008; Rokas, 2008), the relation with structural variation has not been widely explored (Notredame, 2007). This is remarkable given that structural alignments are generally employed to benchmark sequence alignment methods. We demonstrate that in many cases structural alignments can vary dramatically even for small structural changes. Trends observed in the set of crystal structures corroborate those observed in the set of simulation snapshots, albeit alignment differences in the latter set are more pronounced due to larger structural

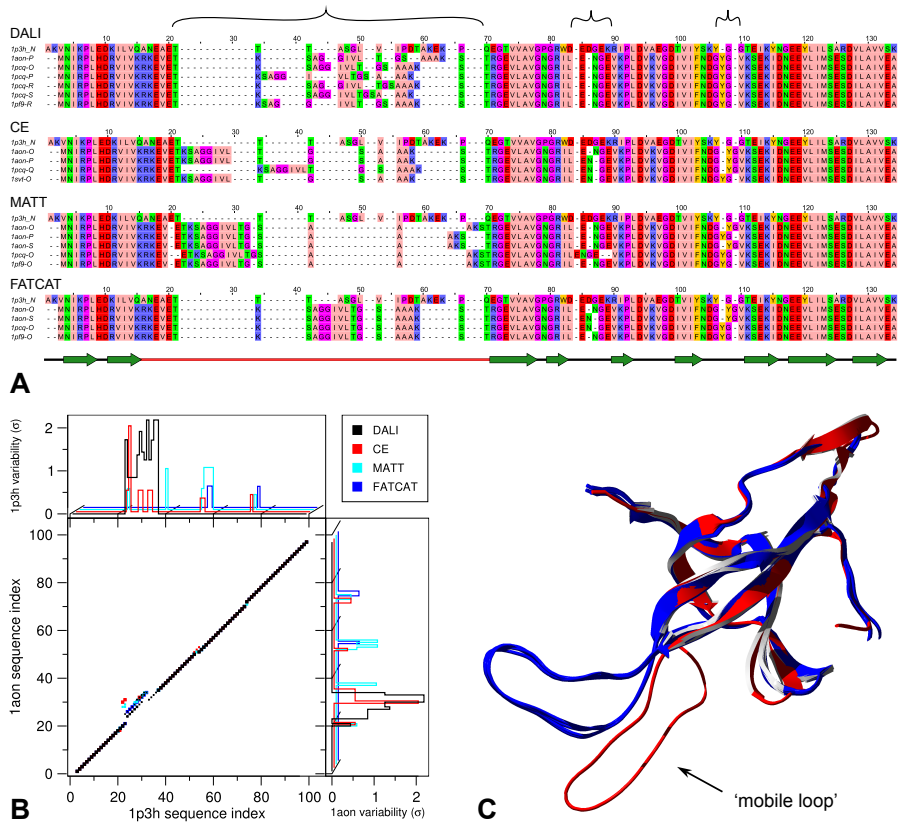


Figure 7.4: An example of the impact of tiny structural variations in the GroES 'mobile loop' that lead to quite dramatic variations in the alignment. **A)** The 'master-slave' alignments with 1p3h-N as master, variable regions marked at the top, and secondary structure with the mobile loop shown in red at the bottom. **B)** Alignment matrix with consistency plots along both axes give an overview of variability in each of the alignments from A). **C)** The different GroEs structures with 1p3h-N in red and 1aon-O and alternatives in blue. Alignment image created using JalView (Clamp et al., 2004) with 'Zappo' colouring; secondary structure assignment according to Xu et al. (1997). Protein structure rendered using SwissPDBViewer (Kaplan and Littlejohn, 2001) and PovRay (`www.povray.org`).

variations.

7.2.2 A depository for alignment variability

It is questionable whether a single reference alignment captures the full width of naturally occurring sequence variability (Godzik, 1996). Yet, current visualization and alignment methods are not designed to take variable regions into account, and they are typically ignored in sequence alignment benchmark protocols. Since variable regions are often important structurally and/or functionally, new approaches for visualization, alignment and benchmarking are desirable.

To this end we have constructed a database of 'flexible' reference alignments. This database is available online `www.ibi.vu.nl/databases/stralivari` and contains all structures and alignments used in this study. For each alignment in our database, variation is visualized using alignment matrices and consistency plots as shown in Figure 7.4B. In addition the database contains the ensemble 'master-slave' alignments as shown in Figure 7.4A. This pinpoints alignment regions that are affected by variability.

7.3 Conclusion

Structural variation, as presented here by alternative crystal structures and molecular dynamics simulations, has a profound effect on structural alignment. The sensitivity to structural variation is a bottleneck for the effective application of structural alignment approaches. This undermines the current basis of all sequence alignment methodologies and is an underestimated problem for the homology assessment used in structural and functional classification. The GroES 'mobile loop' example demonstrates how functionally essential protein regions can coincide with variable structural alignment segments. Our database should therefore be useful for alignment verification and delineation of functionally important protein regions.

7.4 Methods

The HOMSTRAD database of homologous structure alignments (Mizuguchi et al., 1998) was used as a source to select homologous proteins with known structure. HOMSTRAD families containing two homologous proteins (A and B in Figure 7.2) were selected. The corresponding structures were retrieved from the PDB (Berman et al., 2000) and taken as reference. For each reference structure, after equilibration, molecular dynamics simulations were performed for up to 10 ns, and snapshot structures were stored every 1 ns. Standard solvated conditions in the Gromos 43a1 forcefield (Hünenberger et al., 1995) and the Gromacs simulation package (Lindahl et al., 2001) were used (details summarized in additional file 2, Section 7.6). In addition, for each reference structure, we retrieved all alternative PDB structures with

100% sequence identity. In the subsequent analysis only the residues corresponding to the HOMSTRAD sequences were used.

From each pair of reference HOMSTRAD structures, we constructed reference alignments with the widely used structural alignment tool DALI (Holm and Park, 2000). We also used DALI to create pairwise alignments between each reference structure and the alternatives of the other reference structure (PDB and snapshots). The sequence differences between the alignments were calculated using Sum-of-Pairs (SP) scoring implemented in the BAliBASE alignment comparison tool (Thompson et al., 1999). SP scores range from 0 (non-identical) to 1 (identical sequence alignments). Finally we calculated the root mean square deviation (RMSD) between the C_α atoms of the alternative structures and their reference structure using the McLachlan algorithm (McLachlan, 1982) as implemented in the program ProFit version 2.5.3 (`www.bioinf.org.uk/software/profit`).

Our final database consists of 496 proteins (divided over 341 families) for which 3309 snapshot structures could be made and 565 proteins (divided over 395 families) for which we found in total 2998 alternative crystal structures with redundant sequences. A full list of all aligned structures and relevant details is provided in additional file 3 (Section 7.6).

7.5 Authors‘ contributions

All authors designed the research, analyzed the results and wrote the paper. WP and KAF performed the research. All authors read and approved the final manuscript.

7.6 Additional material

Additional file 1
Figure S1: Combined effects of structural variation and sequence variation on the alignment.
This file can be found at
[`www.biomedcentral.com/content/supplementary/1471-2105-9-556-s1.pdf`].

Additional file 2
Table S1: Molecular Dynamics simulation set-up.
This file can be found at
[`www.biomedcentral.com/content/supplementary/1471-2105-9-556-s2.pdf`].

Additional file 3
Table S2: Details of aligned Crystal Structures (a) and Simulation Snapshots (b).
This file can be found at
[`www.biomedcentral.com/content/supplementary/1471-2105-9-556-S3.pdf`].

7.7 Acknowledgements

We like to thank Sander W. Timmer and Anneke van der Reijden for development of the data-analysis scripts and Bernd W. Brandt for the set of redundant protein structures. Financial support was provided by the Netherlands Bioinformatics Centre, BioRange Bioinformatics research programmes SP 3.2.2 and SP 2.3.1.

CHAPTER 8

Structure and function analysis of flexible alignment regions in proteins

Published as:

Pirovano, W., van der Reijden, A., Feenstra, K.A., and Heringa, J. (2009).
Structure and function analysis of flexible alignment regions in proteins.
BMC Bioinformatics, 10(Suppl 13):P6.

This article is part of the supplement: *Highlights from the Fifth International Society for Computational Biology (ISCB) Student Council Symposium, Stockholm, Sweden.*

8.1 Background

Protein structural alignment plays a key role in defining gold standards for a variety of bioinformatics applications. These include homology assessment, phylogenetic tree construction and multiple sequence alignment evaluation. Our recent findings (Pirovano et al., 2008a) however showed that superposition methods are rather sensitive to structural variation. To sidestep the problem of alignment variability, golden standards are often derived from the more conserved and trusted regions. It therefore remains unclear which structural elements characterize alignment variability and what functional information these discarded flexible regions entail.

8.2 Methods

The dataset was taken from Pirovano et al. (2008a) and consists of 565 proteins for which in total 2998 alternative crystal structures with redundant sequences were available. Structural alignments were made using DALI (Holm and Park, 2000).

Alignment variability is defined by the standard deviation of residue shifts over an ensemble of alternate alignments (sigma):

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2}{n}} \tag{8.1}$$

where $\bar{X}$ is the mean of all residue shifts X_i and n is the number of alternate alignments.

The fractions of structural elements are defined as the number of residues involved in a secondary structure element (helix, strand or coil) divided by the total number of residues. Fractions of PROSITE functional classes are defined as the number of residues involved in a functional class by the total number of residues.

8.3 Results

In this study we shed more light on the structural features and functional importance associated with flexible alignment regions. We observe that helices and coils constitute the main source of alignment variability (around 60% and 30%, respectively), while strands appear to be more robust (see Figure 8.1A). Additional alignment inspection shows that many secondary structure elements are not consistently aligned thus giving rise to mismatches between secondary structure types. Functional investigation using Prosite (de Castro et al., 2006) reveals that roughly 20% of all flexible alignment positions correspond to functional sites (see Figure 8.1B), similar to stably aligned regions. Interestingly, post-translational modification sites are strongly represented and particularly phosphorylation sites are prominent. It is therefore unwarranted to assume that these flexible regions only play a minor role in protein function. An

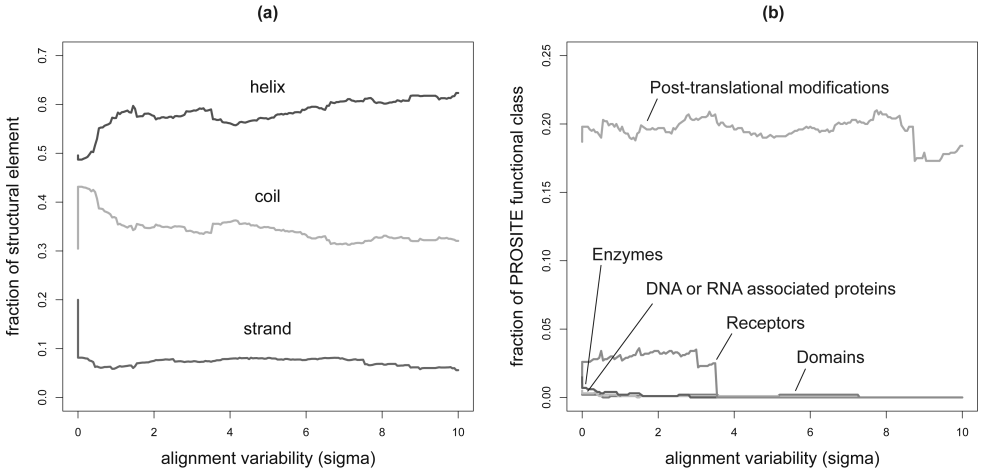


Figure 8.1: Structure and function analysis of alignment variability by means of the standard deviation of residue shifts over an ensemble of alternate alignments (sigma). **A)** Shows the fraction of helical residues increasing with higher variability at the cost of the fraction of coil residues. **B)** Shows the even distribution of functional sites (as grouped by Prosite) over a large range of variability.

example of how the alignment of structural motifs can be impacted by tiny structural variations is given by Figure Figure 8.2, which shows the alignment between a Glutaminyl-tRNA synthetase and a Caspase-8.

8.4 Conclusion

Our results imply that the current 'gold' standard status of structural alignment should be considered 'silver'. Particularly our observation that helices are associated with flexible alignment regions is at odds with currently used alignment strategies. Moreover, given that functional importance is spread evenly between stably and flexibly aligned regions, we conclude that flexible regions cannot be excluded from analysis of functionality in proteins. In order to explore new strategies for homology detection, phylogeny and alignment we propose that, as a first step, more golden standards be developed that can more comprehensively represent the structural, functional and evolutionary signals.

8.5 Acknowledgements

Financial support for this project was provided by the Netherlands Bioinformatics Centre, BioRange Bioinformatics research programme SP 3.2.2.

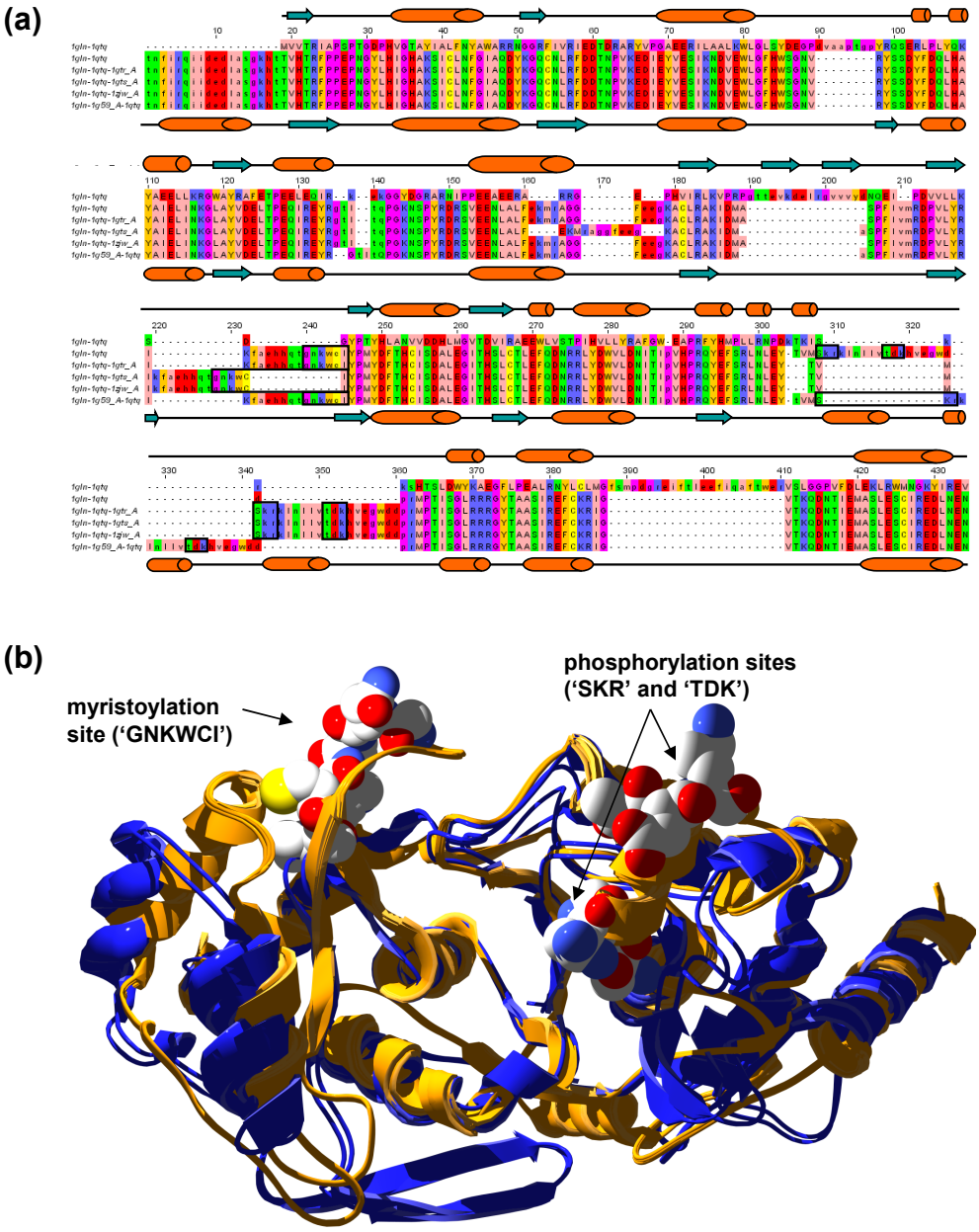


Figure 8.2: Dramatic variations in the alignment between a Glutaminyl-tRNA synthetase (1qtq) and a Caspase-8 (1qtn). For 1qtq, the flexible alignment regions contain two PKC phosphorylation sites ('SKR' and 'TDK') in a helix and a myristoylation site ('GNKWCI') in a coil region. **A)** The 'master-slave' alignments with 1qtn as master, secondary structures at the top (1qtn) and the bottom (1qtq), functional sites marked in apposite windows. **B)** The different Glutaminyl-tRNA synthetase and Caspase-8 structures with 1qtq and alternatives in orange and 1qtn and alternatives in blue. Functional motifs are marked. Image rendered using SwissPDBViewer (Kaplan and Littlejohn, 2001) and PovRay (`www.povray.org`).

CHAPTER 9

Summarizing Discussion

9.1 What is the point of alignment method development?

This thesis describes a four-year journey devoted to protein multiple sequence alignment. It includes the presentation of several automated strategies to enhance both alignment construction and analysis. The large attention that is paid to this field over the past decades is mainly justified by its great contribution to function analyses, even though the complexity of the information stored in DNA and proteins has not nearly been unraveled. In this light the assembly of proper multiple alignments has proved essential in sequence analysis despite the fact that even the most successful alignment algorithms can still fail to provide reliable answers (Pirovano and Heringa, 2008).

From various perspectives there is clearly room for method improvement. A first straightforward indication is given by results obtained on alignment benchmarks. Methods construct better and better alignments over the years but especially in divergent sequence sets difficulties arise. In particular if the sequence identity (percentage of identical residues) shared between sequences drops below 35%, methods do misalign a significant number of residues. It should be kept in mind that structural correspondence between two proteins cannot be guaranteed in the twilight zone of 20-35% sequence identity (Rost, 1999). Nonetheless there are still many examples of homologous relationships with less than 15% sequence identity (Pascarella and Argos, 1992). On the other hand, examples have been found of proteins sharing around 50% sequence identity with a different structure and function (Rost, 1999).

Another very important development has been the advance in terms of speed, making it possible to align hundreds or thousands of sequences in a relatively short amount of time. It must be stressed though that in general speed improvements neg-

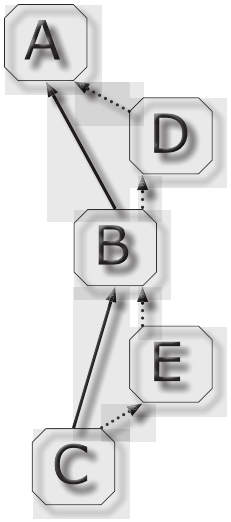


Figure 9.1: Abstract representation of stepping stone alignment. The route taken to align sequence A and C would consist of different steps if solely sequence B or also D and E are included. More intermediate steps (sequences) lead to easier to take distances (alignments).

atively influence the alignment quality. In the advent of Next-Generation Sequencing this becomes even more an issue (see below).

Main difficulties however arise from inherent fundamental issues in alignment construction. A first intrinsic problem is that the evolution of each protein family, or each sequence contained in it, is unique and is determined by specific genetic and environmental signals. This means that also the alignment of a particular sequence set should ideally be guided by specific evolutionary rules. Given that there are thousands of protein families, an automated routine that applies 'general rules' (and for most methods this is the case) is limited in its search for optimal biological solutions. Moreover, biologists often do not realize that actually they are the key players in guiding the process themselves, being the specialists of a particular protein family. In contrast, in fully automated routines no expert knowledge on the particular family is used to steer the alignment.

A second issue concerns the loss of information: we usually do not have the ancestral sequences which would complete the evolutionary picture. An early attempt to overcome this was proposed by Hogeweg and Hesper (1984), who constructed 'internode' (ancestral) sequences that were added to the query set. However, whereas expert knowledge of protein structure or function is based upon experimental evidence, the determination of *e.g.* evolutionary speed and pressure can only be performed by approximation. We therefore have to rely on present-day sequences and should pay attention that our input sufficiently covers the sequence space of the particular fam-

ily. Especially the progressive alignment algorithm, that is used by nearly methods, is sensitive to this issue. An abstract illustration of so-called 'stepping stone alignment' is given in Figure 9.1. The point here is that a multiple alignment routine may align three sequences A, B and C in a different manner given the presence or absence of two other sequences D and E. Moreover the alignment of either sequences A, B and C or A, B and three identical sequences C–1, C–2 and C–3 may already give different alignments between A and B.

Some methods have attempted to include sequence weighting to correct for this, in most cases by upweighting distant sequences as they carry more information about the divergence of the query set. Although in principle this is a good idea, it is only warranted if the alignment is evaluated for correctness. When such a scenario is used in progressive multiple alignment, the increased chance of alignment error for distant sequences often leads to the amplification of misinformation. To quantify this effect, Vogt et al. (1995) compared pairwise alignments with a data bank of structure-based alignments (Pascarella and Argos, 1992) as a standard of truth. To ensure good alignment quality, they included a large set of substitution matrices with optimised gap-penalties. The best scoring combination of global alignment with the Gonnet residue exchange matrix (Gonnet et al., 1992) showed a dramatic increase of alignment error towards in the more distant cases. For example, the results indicated 15% incorrect residue matching on average when sequences with 30% residue identity were aligned. The error rate quickly increased to 45% incorrect matches at 20% residue identity of the aligned sequences, and to 73% error at 15% sequence identity. These statistics clearly show the risk of upweighting the importance of more distance sequence matches. In the next section we therefore further discuss our alternative approach to sequence weighting aimed to better describe the sequence space.

An additional important matter relating to the loss of information, concerns the idea of mapping sequence dissimilarity to evolutionary distance in general. Alignment programs commonly count the the number of substitutions in (pairwise) alignments to estimate the evolutionary distance between (two) sequences (*i.e.* sequence identity). This assumption however is wrong since the number of substitutions observed usually does not reflects the true number of substitutions occurred over time. For instance, an the replacement of an Alanine by a Isoleucine might have undergone several intermediate steps, such as Alanine - Valine - Leucine - Isoleucine. Similarly the apparent conservation of an Alanine might have been the result of a so-called 'back mutation', such as Alanine - Valine - Alanine. Another scenario for the observed conservation of *e.g.* Valines could also be the parallel independent mutation of two Alanines into two Valines due to convergent evolution. As a consequence, the evolutionary effects which can not be captured in alignments lead to the incorrect phylogenetic or guide tree construction. In an attempt to correct for multiple substitutions and convergence, different evolutionary models have been proposed including the Jukes-Cantor and Kimura nucleotide substitution models .

Another inherent difficulty, extensively described in Chapter 7 (Pirovano et al., 2008a) and 8 (Pirovano et al., 2009b), is that alignments are static and do not capture protein dynamics. At a sequence level it is evident that a dynamic display of

alignments would be difficult to read, but also at a structure level the development of dynamic alignment representations is not straightforward. Currently we can only capture structure dynamics in a limited way through the display of different protein conformations in one overview. In other words, even a biologically reliable multiple sequence alignment can only provide an 'average description' of protein dynamics. Alignment ambiguity can easily occur when a protein changes it conformation, for instance to bind to a partner. It can be easily imagined that this ambiguity might have a rather strong impact on subsequent site-specific alignment analysis. In addition it should be stressed that multiple alignments attempt to describe the evolution of single residues, although protein function and structure are in many cases determined by ensembles of residues.

9.2 What is the point of PRALINE?

The bioinformatics textbook *Essential Bioinformatics* by Jin Xiong concludes a section on PRALINE by stating "PRALINE is perhaps the most sophisticated alignment program available. Because of the high complexity of the algorithm, its obvious drawback is the extremely slow computation" (Xiong, 2006). PRALINE allows the integration of additional knowledge gained from external resources. Our major drive has been to create an alignment suite that:

1. is guided by additional evolutionary signals supplementing primary input sequence information;
2. has a general applicability;
3. can optionally provide *ad hoc* solutions for particular alignment cases;
4. and is easy to use.

The main result of this focus is that PRALINE provides some important strategies to address the inherent alignment problems described above. The re-use of sequence information through local and global pre-processing (Heringa, 1999, 2002) and the inclusion of homologous sequence data retrieved through PSI-BLAST (Simossis et al., 2005) were initial steps to better describe the specific sequence space and evolutionary history of a protein family. To some degree the evolutionary gap between the input sequences becomes filled, optimising the chance of more accurate 'stepping stone' alignments. In contrast to methods that upweight the information of distant sequences (sequence weighting), PRALINE's pre-processing and PSI-BLAST protocols tend to exclude remote sequence signals from the pre-profiles. As a results the intermediate steps of a PRALINE alignment, being closer to each other, are safer to take whereas sequence weighting might lead to slippage when jumping to more distant stones.

The extension of the protocol with predicted secondary (Pirovano et al., 2009a) and transmembrane structure information (Pirovano et al., 2008b) is even more relevant to

stepping stone alignments. The approach exploits the 'structure-is-more-conserved-than-sequence' principle by blending in structural information that effectively leads to reduced sequence distances. Another advantage is that structure predictions implicitly take sequence context into account, *i.e.* the prediction of a helix is highly dependent on patterns in neighbouring residues. As a consequence the alignment of functional sites, which tend to cluster in structure space, will be enhanced.

With the PRALINE^{TM} routine (Chapter 3; Pirovano et al., 2008a) we provide a suitable example of how specific sequence sets, such as transmembrane proteins, can be better aligned using tailored evolutionary schemes. Different families are the result of context-dependent evolutionary processes, and alignment methods should take this into account. This is beautifully illustrated by our observation that amino acid patterns differ between transmembrane regions and soluble regions, but also that there are compositional differences between soluble proteins and soluble regions of transmembrane proteins. Apparently, sequence context is a central issue here. The importance of membrane proteins should not be underestimated as this class of proteins is largely represented in sequence databases and often assume essential roles as receptors. Given that these receptors are also regularly implicated in diseases such as cancer, it is perhaps rather surprising that to date PRALINE is the only available method that implements specific technology to align transmembrane proteins.

Future alignment work will include the implementation of other sources of family-specific knowledge, such as:

1. *Motifs*: Short conserved stretches within a protein family could function as anchors of the multiple alignment.

2. *Solvent accessibility*: Differences in the evolutionary processes behind surface and core residues can be taken into account.

3. *Protein networks*: Information about interacting partners and their conservation through out other species can help to guide the alignment.

A major advantage of these information sources is that most of them can be reliably predicted from sequence information alone and would thus augment the general applicability of the tool set.

9.3 A note on structure prediction reliability

In the abstract of Chapter 4 the following statement is made: 'The evolutionary advantage of using predicted structure information outweighs the chances of misprediction'. As mentioned before, the structural information we have incorporated in our alignment program PRALINE relies on the basic concept that a protein's structure is more conserved than it's sequence. It could be questioned however whether, given that sequence alignment should benefit from structural guidance, prediction reliabilities are sufficient enough. Indeed the success of structure-guided alignment methods is dependent on the quality of the predictions.

The current state-of-the-art sustained secondary structure prediction accuracy is around 80%, as for instance attained by the Porter method (Pollastri and McLysaght, 2005; for a review see Pirovano and Heringa, 2009). For transmembrane topology prediction similar accuracies have been reported (Jones, 2007). The transition accuracy from which point alignment methods can benefit from their predictions is around 65% for secondary structure (Rost, 2009). This means that despite the room for improvement of prediction tools, sequence alignment can already largely take advantage of them. Finally, structure prediction tools are 'bounded by the intrinsic ambiguity of mapping three-dimensional atom coordinates into secondary structure classes' (Pollastri and McLysaght, 2005). It is therefore questionable whether future prediction methods will actually ever yield optimal results and consequently what the upper limit is for alignment improvement.

A crucial factor in the prediction of secondary structure is paradoxically the quality of the input multiple sequence alignment. In fact, where 15 years of development of secondary structure prediction techniques has led to an increase of about 7% in prediction accuracy, alternative alignments obtained from different alignment programs lead to secondary structure prediction accuracies varying easily over 20% (Heringa, 2000b). For optimal alignment it is important to carefully select an appropriate set of orthologous sequences that can be trusted to fold into the same secondary structural elements. It should be stressed that even if an optimal set of orthologous sequences is assembled and aligned exactly according to their evolutionary relationships, slight differences in the length of secondary structure elements in the various orthologous sequences will lead to an alignment where the flanking regions of matched helices and strands will be uneven, resulting in noise that will negatively affect the delineation of the secondary structure elements of the query sequence. The intricate evolutionary relationships between orthologous sequences and associated structural variation is believed to be a major burden for optimal prediction based on multiple alignment information.

As far as the improvement of alignment quality itself is concerned, previous investigations have shown that the use of DSSP-assigned 'true' secondary structure information does not necessarily lead to better alignments (Przybylski and Rost, 2004; Simossis, 2005). Analogous observations based on the development of fold recognition methods led to the the conclusion that secondary structure-aware alignment yields better results when the secondary structure is predicted for both the target and the template is actually better rather than predicted on the target and observed (DSSP) on the template (Przybylski and Rost, 2004). Apparently the systematic errors made by prediction tools are less harmful than possible inconsistencies with assigned elements from crystal structures.

In our opinion another essential step towards a better structural integration, is the selection (or eventually creation) of a proper substitution matrix. The quality of the data used to build the matrix appears to be more important than the amount of data used: the PHAT and Lüthy matrices were build from solely 598 and 522 sequences, respectively. We however think that it would be interesting to study whether a secondary structure matrix based on a non-redundant PDB database yields

better results.

9.4 Alignments are compositions of functional harmony ... and disharmony, but no atonality

Further aspects of a rather surprising alignment 'vivacity' were highlighted during the Sequence Harmony project (Pirovano et al., 2006; Feenstra et al., 2007). In many cases we judge alignments based on their degree of conservation: conserved regions are often strong indicators for functional importance and also give us the impression that our alignment is reliable. As a consequence, conserved signals are at the core of various sequence analysis applications such as structure prediction and phylogenetic tree construction. Less conserved regions also have many stories to tell us but usually receive much less attention due to their complex interpretation. A striking example that underscores this issue is given by multiple alignment benchmarks, which often discard uncertain or unreliable alignment regions from testing.

Also in the case of functional specificity determination most methods give a high priority to within-class conservation. A thorough literature study on the Smad family of transcription factors, however, revealed that there are also many unconserved specificity determining sites (Pirovano et al., 2006). Other families tested, such as the Ras-superfamily of GTPases, fully confirm this finding. As a consequence, standard approaches based on within-class conservation may miss many specificity determining sites. The Sequence Harmony method instead captures *residue differences* between subfamilies and therefore selects 'disharmonious' sites, *i.e.* classes showing different amino acid compositions, while disregarding conservation.

In traditional sequence alignment there is a large emphasis on positional conservation, while the Sequence Harmony approach shows that functional importance may well reside in apparently unconserved regions. In any case there is no 'atonality', since in all protein families studied there are distinct regions and sites that carry more specificity signals than others.

9.5 Sequence context: know thy neighbours

A second important observation made during the Sequence Harmony project was that residues should indeed not be viewed as isolated individuals, but rather as 'social groups' that act together in carrying out their functions. In other words, specificity determining residues can be more accurately predicted if we take their context into account. Sequence context can be easily scanned for specificity determining residues that are surrounded by sequentially close neighbours with the same property. The principle that neighbouring residues reveal stronger signals is certainly not new: *i.e.* for many years secondary structure prediction tools have relied on sequence residue context, but in other fields the importance of sequence context is still underestimated. Sequence Harmony is one of the first methods to implement sequence context into its

functional specificity prediction algorithm.

Moreover, knowledge about structural context can be even more useful, as function is associated with the three-dimensional protein structure rather than its sequence. However, this requires tertiary structural knowledge, which is not always available, and additional criteria to properly assess spatial clustering. Our study shows that the availability of a tertiary structure for a single member of the alignment can already be of great help to assign functional clusters.

Taking a step back to alignment algorithms, the choice to treat amino acids as single entities here is arguable as well. At present, methods compare profile columns rather than profile regions although it is clear that it is crucial to also take into account structural and functional contexts. Chapter 7 (Pirovano and Heringa, 2008) and 8 (Pirovano et al., 2009b) display the limitations of static alignments: small structural protein movements can induce large variations in alignments and make it unfeasible to asses one-to-one functional equivalence between residues from different proteins. Modern alignment benchmarks should attempt to enlarge their horizons and find solutions to 'ambiguous' alignment regions. Finally also new alignment visualization strategies should be designed that are capable of mapping equivalent regions, rather than just static residue correspondence.

9.6 Alignment issues in next-generation sequencing data

This thesis has a clear focus on multiple alignment of protein sequences and knowledge that can be derived from it. To some extent the alignment of DNA sequences shows similar aspects. For instance, an alignment approach for two DNA segments may still follow the dynamic programming protocol using a nucleotide substitution matrix, although evolutionary operations such as inversions and repeats (or transpositions) are not easily implemented using the protocol. In the advent of Next-Generation Sequencing (NGS) experiments there is however a rapidly growing attention for whole-genome comparison, for which the dynamic programming algorithm is far too slow NGS data analysis is still in an initial phase and many upcoming (joint) projects will focus on improving genome assembly, data-analysis pipelines, user-interfaces and data-storage.

For whole-genome alignment recently some tools have become available that implement extremely fast alignment algorithms (Li et al., 2008a; Li et al., 2008b; Langmead et al., 2009) but at the cost of alignment accuracy. Indeed the computational limitations have restricted tools to align genomes by mapping short stretches of query DNA (reads) to a reference genome based on sequence identity only. This is in high contrast with alignment tools such as PRALINE where the alignment follows an elaborate evolutionary scheme of substitution exchange matrices and gap penalties. In fact it is almost a paradoxical observation that after many years of alignment optimization, genome alignment is currently reduced to 'blind mapping' of reads without using any kind of evolutionary information. Another drawback of mapping such short fragments is that these can not use signals embedded in neighbouring segments.

It can be expected that the knowledge gathered from multiple alignment development as described in this thesis will also be useful for advances in whole-genome alignment where the versatility of the alignment problems is even greater. Further benefits should come from the integration of metadata into genome alignment routines, such as cell context, gene regulatory and signaling networks, protein-protein interactions and expert knowledge. The incorporation of additional information will help to bridge the gap between different genome sequences, so that their evolutionary and functional relationships can be more reliably assessed.

In the advent of the 1000 dollar genome, screening individual genetic variance will also open the gateway to personalized medicine. The identification of single base mutations (or SNPs) and insertions or deletions through NGS might unravel the genetic bases of many complex diseases. This might consequently lead to tailored medical solutions given a person's particular DNA profile. Multiple alignment and other sequence analysis tools are likely to be among the methods of choice in achieving these goals and deciphering the mechanisms behind one of the vital issues in life: our health and quality of life.

References

Abagyan, R. A. and Batalov, S. (1997). Do aligned sequences share the same fold? *J Mol Biol*, 273(1):355–368.

Affourtit, C., Albury, M. S., Crichton, P. G., and Moore, A. L. (2002). Exploring the molecular nature of alternative oxidase regulation and catalysis. *FEBS Lett*, 510(3):121–126.

Altschul, S. F. (1998). Generalized affine gap costs for protein sequence alignment. *Proteins*, 32(1):88–96.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

Altschul, S. F., Madden, T. L., Schffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402.

Andersson, M. E. and Nordlund, P. (1999). A revised model of the active site of alternative oxidase. *FEBS Lett*, 449(1):17–22.

Attwood, T. K., Beck, M. E., Bleasby, A. J., Degtyarenko, K., Michie, A. D., and Parry-Smith, D. J. (1997). Novel developments with the prints protein fingerprint database. *Nucleic Acids Res*, 25(1):212–217.

Bahr, A., Thompson, J. D., Thierry, J. C., and Poch, O. (2001). Balibase (benchmark alignment database): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res*, 29(1):323–326.

Baldi, P., Brunak, S., Frasconi, P., Soda, G., and Pollastri, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15(11):937–946.

Bauer, B., Mirey, G., Vetter, I. R., García-Ranea, J. A., Valencia, A., Wittinghofer, A., Camonis, J. H., and Cool, R. H. (1999). Effector recognition by the small gtp-binding proteins ras and ral. *J Biol Chem*, 274(25):17763–17770.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Res*, 28(1):235–242.

Bhaduri, A., Pugalenthi, G., and Sowdhamini, R. (2004). Pass2: an automated database of protein alignments organised as structural superfamilies. *BMC Bioinformatics*, 5:35.

Bucher, P., Karplus, K., Moeri, N., and Hofmann, K. (1996). A flexible motif search technique based on generalized profiles. *Comput Chem*, 20(1):3–23.

Capriotti, E., Fariselli, P., Rossi, I., and Casadio, R. (2004). A shannon entropy-based filter detects high- quality profile-profile alignments in searches for remote homologues. *Proteins*, 54(2):351–360.

Carrillo, H. and Lipman, D. (1988). The multiple sequence alignment problem in biology. *SIAM J. Appl. Math.*, 48(5):1073–1082.

Chen, Y. G., Hata, A., Lo, R. S., Wotton, D., Shi, Y., Pavletich, N., and Massagué, J. (1998). Determinants of specificity in tgf-beta signal transduction. *Genes Dev*, 12(14):2144–2152.

Chen, Y. G. and Massagué, J. (1999). Smad1 recognition and activation by the alk1 group of transforming growth factor-beta family receptors. *J Biol Chem*, 274(6):3672–3677.

Clamp, M., Cuff, J., Searle, S. M., and Barton, G. J. (2004). The jalview java alignment editor. *Bioinformatics*, 20(3):426–427.

Considine, M. J., Holtzapffel, R. C., Day, D. A., Whelan, J., and Millar, A. H. (2002). Molecular distinction between alternative oxidase from monocots and dicots. *Plant Physiol*, 129(3):949–953.

Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). A model for evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, 5:345–352.

de Castro, E., Sigrist, C. J. A., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., Bairoch, A., and Hulo, N. (2006). Scanprosite: detection of prosite signature matches and prorule-associated functional and structural residues in proteins. *Nucleic Acids Res*, 34(Web Server issue):W362–W365.

del Sol Mesa, A., Pazos, F., and Valencia, A. (2003). Automatic methods for predicting functionally important residues. *J Mol Biol*, 326(4):1289–1302.

Do, C. B., Mahabhashyam, M. S. P., Brudno, M., and Batzoglou, S. (2005). Probcons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*, 15(2):330–340.

Donald, J. E. and Shakhnovich, E. I. (2005). Predicting specificity-determining residues in two large eukaryotic transcription factor families. *Nucleic Acids Res*, 33(14):4455–4465.

Edgar, R. C. (2004a). Local homology recognition and distance measures in linear time using compressed amino acid alphabets. *Nucleic Acids Res*, 32(1):380–385.

Edgar, R. C. (2004b). Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113.

Edgar, R. C. (2004c). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–1797.

Edgar, R. C. and Sjölander, K. (2004). Coach: profile-profile alignment of protein families using hidden markov models. *Bioinformatics*, 20(8):1309–1318.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guig, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., and Emmanouil T Dermitzakis, *et al*. (2007). Identification and analysis of functional elements in 1genome by the encode pilot project. *Nature*, 447(7146):799–816.

Feenstra, K. A., Pirovano, W., Krab, K., and Heringa, J. (2007). Sequence harmony: detecting functional specificity from alignments. *Nucleic Acids Res*, 35(Web Server issue):W495–W498.

Feng, D. F. and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol*, 25(4):351–360.

Feng, X.-H. and Derynck, R. (2005). Specificity and versatility in tgf-beta signaling through smads. *Annu Rev Cell Dev Biol*, 21:659–693.

Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. L., and Bateman, A. (2006). Pfam: clans, web tools and services. *Nucleic Acids Res*, 34(Database issue):D247–D251.

Forrest, L. R., Tang, C. L., and Honig, B. (2006). On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophys J*, 91(2):508–517.

Fu, D., Libson, A., Miercke, L. J., Weitzman, C., Nollert, P., Krucinski, J., and Stroud, R. M. (2000). Structure of a glycerol-conducting channel and the basis for its selectivity. *Science*, 290(5491):481–486.

Galtier, N., Gouy, M., and Gautier, C. (1996). Seaview and phylo_win: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci*, 12(6):543–548.

Germain, S., Howell, M., Esslemont, G. M., and Hill, C. S. (2000). Homeodomain and winged-helix transcription factors recruit activated smads to distinct promoter elements via a common smad interaction motif. *Genes Dev*, 14(4):435–451.

Gerstein, M. and Levitt, M. (1998). Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein Sci*, 7(2):445–456.

Gille, C. and Frömmel, C. (2001). Strap: editor for structural alignments of proteins. *Bioinformatics*, 17(4):377–378.

Ginalski, K., Pas, J., Wyrwicz, L. S., von Grotthuss, M., Bujnicki, J. M., and Rychlewski, L. (2003). Orfeus: Detection of distant homology using sequence profiles and predicted secondary structure. *Nucleic Acids Res*, 31(13):3804–3807.

Godzik, A. (1996). The structural alignment between two proteins: is there a unique answer? *Protein Sci*, 5(7):1325–1338.

Gonnet, G. H., Cohen, M. A., and Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database. *Science*, 256(5062):1443–1445.

Gotoh, O. (1995). A weighting system and algorithm for aligning many phylogenetically related sequences. *Comput Appl Biosci*, 11(5):543–551.

Gotoh, O. (1996). Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J Mol Biol*, 264(4):823–838.

Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A*, 84(13):4355–4358.

Guex, N. and Peitsch, M. C. (1997). Swiss-model and the swiss-pdbviewer: an environment for comparative protein modeling. *Electrophoresis*, 18(15):2714–2723.

Hannenhalli, S. S. and Russell, R. B. (2000). Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol*, 303(1):61–76.

Haussler, D., Krogh, A., Mian, I. S., and Sjolander, K. (1993). Protein modeling using hidden markov models: analysis of globins. In *Proc. Proceeding of the Twenty-Sixth Hawaii International Conference on System Sciences*, pages 792–802.

Heger, A., Lappe, M., and Holm, L. (2004). Accurate detection of very sparse sequence motifs. *J Comput Biol*, 11(5):843–857.

Henikoff, J. G. and Henikoff, S. (1996). Using substitution probabilities to improve position-specific scoring matrices. *Comput Appl Biosci*, 12(2):135–143.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919.

Heringa, J. (1999). Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. *Comput Chem*, 23(3-4):341–364.

Heringa, J. (2000a). Computational methods for protein secondary structure prediction using multiple sequence alignments. *Curr Protein Pept Sci*, 1(3):273–301.

Heringa, J. (2000b). Predicting secondary structure from protein sequences. *Bioinformatics: Sequence, structure and databanks*, Higgins, D.G. and Taylor, W.R. eds., Oxford University Press, 113-142.

Heringa, J. (2002). Local weighting schemes for protein multiple sequence alignment. *Comput Chem*, 26(5):459–477.

Heringa, J. and Pirovano, W. (2007). Sequence similarity searches. *Bioinformatics, Method Express Series*, Dear, P. ed., Scion Publishing Ltd, Oxfordshire, UK, 39-67.

Heringa, J. and Taylor, W. R. (1997). Three-dimensional domain duplication, swapping and stealing. *Curr Opin Struct Biol*, 7(3):416–421.

Herraez, A. (2006). Biomolecules in the Computer: Jmol to the Rescue. *Biochem Mol Biol Edu*, 34(4):255.

Hirosawa, M., Totoki, Y., Hoshida, M., and Ishikawa, M. (1995). Comprehensive study on iterative algorithms of multiple sequence alignment. *Comput Appl Biosci*, 11(1):13–18.

Hogeweg, P. and Hesper, B. (1984). The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *J Mol Evol*, 20(2):175–186.

Holm, L., Ouzounis, C., Sander, C., Tuparev, G., and Vriend, G. (1992). A database of protein structure families with common folding motifs. *Protein Sci*, 1(12):1691–1698.

Holm, L. and Park, J. (2000). Dalilite workbench for protein structure comparison. *Bioinformatics*, 16(6):566–567.

Holmes, I. and Durbin, R. (1998). Dynamic programming alignment accuracy. *J Comput Biol*, 5(3):493–504.

Huang, X. and Miller, W. (1991). A time-efficient, linear-space local similarity algorithm. *Adv. Appl. Math.*, 12:337–357.

Hünenberger, P. H., Mark, A. E., and van Gunsteren, W. F. (1995). Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. *J Mol Biol*, 252(4):492–503.

Huse, M., Muir, T. W., Xu, L., Chen, Y. G., Kuriyan, J., and Massagué, J. (2001). The tgf beta receptor activation process: an inhibitor- to substrate-binding switch. *Mol Cell*, 8(3):671–682.

Jaroszewski, L., Rychlewski, L., and Godzik, A. (2000). Improving the quality of twilight-zone alignments. *Protein Sci*, 9(8):1487–1496.

Jones, D. T. (1998). Do transmembrane protein superfolds exist? *FEBS Lett*, 423(3):281–285.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, 292(2):195–202.

Jones, D. T. (2007). Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, 23(5):538–544.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1994a). A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, 33(10):3038–3049.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1994b). A mutation data matrix for transmembrane proteins. *FEBS Lett*, 339(3):269–275.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637.

Kalinina, O. V., Mironov, A. A., Gelfand, M. S., and Rakhmaninova, A. B. (2004). Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci*, 13(2):443–456.

Käll, L., Krogh, A., and Sonnhammer, E. L. L. (2004). A combined transmembrane topology and signal peptide prediction method. *J Mol Biol*, 338(5):1027–1036.

Käll, L., Krogh, A., and Sonnhammer, E. L. L. (2005). An hmm posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics*, 21 Suppl 1:i251–i257.

Kaplan, W. and Littlejohn, T. G. (2001). Swiss-pdb viewer (deep view). *Brief Bioinform*, 2(2):195–197.

Katoh, K., ichi Kuma, K., Toh, H., and Miyata, T. (2005). Mafft version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, 33(2):511–518.

Katoh, K., Misawa, K., ichi Kuma, K., and Miyata, T. (2002). Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res*, 30(14):3059–3066.

Kimura, M. (1979). The neutral theory of molecular evolution. *Sci Am*, 241(5):98–100, 102, 108 passim.

Kolodny, R., Koehl, P., and Levitt, M. (2005). Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J Mol Biol*, 346(4):1173–1188.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J Mol Biol*, 305(3):567–580.

Kuipers, W., Oliveira, L., Vriend, G., and Ijzerman, A. P. (1997). Identification of class-determining residues in g protein-coupled receptors by sequence analysis. *Receptors Channels*, 5(3-4):159–174.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., and M. Doyle, International Human Genome Sequencing Consortium, *et al*. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007). Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–2948.

Li, H., Ruan, J., and Durbin, R. (2008a). Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–1858.

Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008b). Soap: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714.

Li, W.-H. and Graur, D. (1991). *Fundamentals of molecular evolution*. Sinauer, Sunderland, MA.

Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol*, 257(2):342–358.

Lin, K., Simossis, V. A., Taylor, W. R., and Heringa, J. (2005). A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics*, 21(2):152–159.

Lindahl, E., Hess, B., and Spoel, D. (2001). Gromacs 3.0: a package for molecular simulation and trajectory analysis. *J Mol Mod*, 7(8):121–126.

Lindqvist, Y., Huang, W., Schneider, G., and Shanklin, J. (1996). Crystal structure of delta9 stearoyl-acyl carrier protein desaturase from castor seed and its relationship to other di-iron proteins. *EMBO J*, 15(16):4081–4092.

Livingstone, C. D. and Barton, G. J. (1996). Identification of functional residues and secondary structure from protein multiple sequence alignment. *Methods Enzymol*, 266:497–512.

Lo, R. S., Chen, Y. G., Shi, Y., Pavletich, N. P., and Massagué, J. (1998). The l3 loop: a structural motif determining specific interactions between smad proteins and tgf-beta receptors. *EMBO J*, 17(4):996–1005.

Lupyan, D., Leo-Macias, A., and Ortiz, A. R. (2005). A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, 21(15):3255–3263.

Lüthy, R., McLachlan, A. D., and Eisenberg, D. (1991). Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins*, 10(3):229–239.

Madera, M. (2008). Profile comparer: a program for scoring and aligning profile hidden markov models. *Bioinformatics*, 24(22):2630–2631.

Maiorov, V. N. and Crippen, G. M. (1995). Size-independent comparison of protein three-dimensional structures. *Proteins*, 22(3):273–283.

Maizel, J. V. and Lenk, R. P. (1981). Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proc Natl Acad Sci U S A*, 78(12):7665–7669.

Massagué, J., Seoane, J., and Wotton, D. (2005). Smad transcription factors. *Genes Dev*, 19(23):2783–2810.

Mayr, G., Domingues, F. S., and Lackner, P. (2007). Comparative analysis of protein structure alignments. *BMC Struct Biol*, 7:50.

McLachlan, A. (1982). Rapid comparison of protein structures. *Acta Cryst*, A38:871–873.

Menke, M., Berger, B., and Cowen, L. (2008). Matt: local flexibility aids protein multiple structure alignment. *PLoS Comput Biol*, 4(1):e10.

Mihalek, I., Res, I., and Lichtarge, O. (2004). A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J Mol Biol*, 336(5):1265–1282.

Mirny, L. A. and Gelfand, M. S. (2002). Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J Mol Biol*, 321(1):7–20.

Mizuguchi, K., Deane, C. M., Blundell, T. L., and Overington, J. P. (1998). Homstrad: a database of protein structure alignments for homologous families. *Protein Sci*, 7(11):2469–2471.

Mizuide, M., Hara, T., Furuya, T., Takeda, M., Kusanagi, K., Inada, Y., Mori, M., Imamura, T., Miyazawa, K., and Miyazono, K. (2003). Two short segments of smad3 are important for specific interaction of smad3 with c-ski and snon. *J Biol Chem*, 278(1):531–536.

Morgenstern, B. (2004). Dialign: multiple dna and protein sequence alignment at bibiserv. *Nucleic Acids Res*, 32(Web Server issue):W33–W36.

Morgenstern, B., Dress, A., and Werner, T. (1996). Multiple dna and protein sequence alignment based on segment-to-segment comparison. *Proc Natl Acad Sci U S A*, 93(22):12098–12103.

Mosca, R. and Schneider, T. R. (2008). Rapido: a web server for the alignment of protein structures in the presence of conformational changes. *Nucleic Acids Res*, 36(Web Server issue):W42–W46.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453.

Ng, P. C., Henikoff, J. G., and Henikoff, S. (2000). Phat: a transmembrane-specific substitution matrix. predicted hydrophobic and transmembrane. *Bioinformatics*, 16(9):760–766.

Notredame, C. (2007). Recent evolutions of multiple sequence alignment algorithms. *PLoS Comput Biol*, 3(8):e123.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–217.

Obenauer, J. C., Denson, J., Mehta, P. K., Su, X., Mukatira, S., Finkelstein, D. B., Xu, X., Wang, J., Ma, J., Fan, Y., Rakestraw, K. M., Webster, R. G., Hoffmann, E., Krauss, S., Zheng, J., Zhang, Z., and Naeve, C. W. (2006). Large-scale sequence analysis of avian influenza isolates. *Science*, 311(5767):1576–1580.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G., and Notredame, C. (2004). 3dcoffee: combining protein sequences and structures within multiple sequence alignments. *J Mol Biol*, 340(2):385–395.

Parry-Smith, D. J., Payne, A. W., Michie, A. D., and Attwood, T. K. (1998). Cinema–a novel colour interactive editor for multiple alignments. *Gene*, 221(1):GC57–GC63.

Pascarella, S. and Argos, P. (1992). A data bank merging related protein structures and sequences. *Protein Eng*, 5(2):121–137.

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with fastp and fasta. *Methods Enzymol*, 183:63–98.

Pei, J. and Grishin, N. V. (2007). Promals: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, 23(7):802–808.

Pirovano, W., Feenstra, K. A., and Heringa, J. (2006). Sequence comparison by sequence harmony identifies subtype-specific functional sites. *Nucleic Acids Res*, 34(22):6540–6548.

Pirovano, W., Feenstra, K. A., and Heringa, J. (2008a). The meaning of alignment: lessons from structural diversity. *BMC Bioinformatics*, 9(1):556.

Pirovano, W., Feenstra, K. A., and Heringa, J. (2008b). Praline$^{\mathrm{tm}}$: a strategy for improved multiple alignment of transmembrane proteins. *Bioinformatics*, 24(4):492–497.

Pirovano, W. and Heringa, J. (2008). Multiple sequence alignment. *Methods Mol Biol*, 452:143–161.

Pirovano, W. and Heringa, J. (2009). Protein secondary structure prediction. *Methods Mol Biol*, in press.

Pirovano, W., Simossis, V. A., and Heringa, J. (2009a). Secondary structure-guided multiple sequence alignment. *submitted*. in press.

Pirovano, W., van der Reijden, A., Feenstra, K. A., and Heringa, J. (2009b). Structure and function analysis of flexible alignment regions in proteins. *BMC Bioinformatics*, 10(Suppl 13):P6.

Pollastri, G. and McLysaght, A. (2005). Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, 21(8):1719–1720.

Pollastri, G., Przybylski, D., Rost, B., and Baldi, P. (2002). Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, 47(2):228–235.

Przybylski, D. and Rost, B. (2004). Improving fold recognition without folds. *J Mol Biol*, 341(1):255–269.

Punta, M., Forrest, L. R., Bigelow, H., Kernytsky, A., Liu, J., and Rost, B. (2007). Membrane protein prediction methods. *Methods*, 41(4):460–474.

Randall, R. A., Germain, S., Inman, G. J., Bates, P. A., and Hill, C. S. (2002). Different smad2 partners bind a common hydrophobic pocket in smad2 via a defined proline-rich motif. *EMBO J*, 21(1-2):145–156.

Reuther, G. W. and Der, C. J. (2000). The ras branch of small gtpases: Ras family members don't fall far from the tree. *Curr Opin Cell Biol*, 12(2):157–165.

Rokas, A. (2008). Genomics. lining up to avoid bias. *Science*, 319(5862):416–417.

Rost (2009). Personal communication.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng*, 12(2):85–94.

Rost, B. and Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol*, 232(2):584–599.

Rychlewski, L., Jaroszewski, L., Li, W., and Godzik, A. (2000). Comparison of sequence profiles. strategies for structural predictions using sequence information. *Protein Sci*, 9(2):232–241.

Sadreyev, R. and Grishin, N. (2003). Compass: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol*, 326(1):317–336.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425.

Sali, A. and Blundell, T. L. (1990). Definition of general topological equivalence in protein structures. a procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol*, 212(2):403–428.

Sammeth, M. and Heringa, J. (2006). Global multiple-sequence alignment with repeats. *Proteins*, 64(1):263–274.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M., and Smith, M. (1977). Nucleotide sequence of bacteriophage phi x174 dna. *Nature*, 265(5596):687–695.

Service, R. F. (2006). Gene sequencing. the race for the $1000 genome. *Science*, 311(5767):1544–1546.

Shafrir, Y. and Guy, H. R. (2004). Stam: simple transmembrane alignment method. *Bioinformatics*, 20(5):758–769.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech J*, 27:379–423, 623–656.

Shatsky, M., Nussinov, R., and Wolfson, H. J. (2004). A method for simultaneous alignment of multiple protein structures. *Proteins*, 56(1):143–156.

Shenkin, P. S., Erman, B., and Mastrandrea, L. D. (1991). Information-theoretical entropy as a measure of sequence variability. *Proteins*, 11(4):297–313.

Shi, J., Blundell, T. L., and Mizuguchi, K. (2001). Fugue: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol*, 310(1):243–257.

Shindyalov, I. N. and Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Eng*, 11(9):739–747.

Simossis, V. (2005). *From sequence to structure and back again: an alignment tale.* PhD thesis, VU University Amsterdam.

Simossis, V. A. and Heringa, J. (2003). The praline online server: optimising progressive multiple alignment on the web. *Comput Biol Chem*, 27(4-5):511–519.

Simossis, V. A. and Heringa, J. (2005). Praline: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res*, 33(Web Server issue):W289–W294.

Simossis, V. A., Kleinjung, J., and Heringa, J. (2005). Homology-extended sequence alignment. *Nucleic Acids Res*, 33(3):816–824.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197.

Söding, J. (2005). Protein homology detection by hmm-hmm comparison. *Bioinformatics*, 21(7):951–960.

Sonnhammer, E. L., von Heijne, G., and Krogh, A. (1998). A hidden markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol*, 6:175–182.

Stenmark, H. and Olkkonen, V. M. (2001). The rab gtpase family. *Genome Biol*, 2(5):REVIEWS3007.

Stenmark, H., Valencia, A., Martinez, O., Ullrich, O., Goud, B., and Zerial, M. (1994). Distinct structural elements of rab5 define its functional specificity. *EMBO J*, 13(3):575–583.

Stoye, J., Moulton, V., and Dress, A. W. (1997). Dca: an efficient implementation of the divide-and-conquer approach to simultaneous multiple sequence alignment. *Comput Appl Biosci*, 13(6):625–626.

Taylor, W. R. and Orengo, C. A. (1989). Protein structure alignment. *J Mol Biol*, 208(1):1–22.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–4680.

Thompson, J. D., Koehl, P., Ripp, R., and Poch, O. (2005). Balibase 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, 61(1):127–136.

Thompson, J. D., Plewniak, F., and Poch, O. (1999). Balibase: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, 15(1):87–88.

Tomii, K. and Akiyama, Y. (2004). Forte: a profile-profile comparison tool for protein fold recognition. *Bioinformatics*, 20(4):594–595.

Tusnády, G. E., Dosztányi, Z., and Simon, I. (2005). Pdb_tm: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res*, 33(Database issue):D275–D278.

Tusnády, G. E. and Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol*, 283(2):489–506.

Tusnády, G. E. and Simon, I. (2001). The hmmtop transmembrane topology prediction server. *Bioinformatics*, 17(9):849–850.

Vanlerberghe, G. C. and McIntosh, L. (1997). Alternative oxidase: From gene to function. *Annu Rev Plant Physiol Plant Mol Biol*, 48:703–734.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., and C. A. Evans, *et al.* (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

Vogt, G., Etzold, T., and Argos, P. (1995). An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *J Mol Biol*, 249(4):816–831.

von Ohsen, N., Sommer, I., Zimmer, R., and Lengauer, T. (2004). Arby: automatic protein structure prediction using profile-profile alignment and confidence measures. *Bioinformatics*, 20(14):2228–2235.

Wagner, A. M. and Moore, A. L. (1997). Structure and function of the plant alternative oxidase: its putative role in the oxygen defence mechanism. *Biosci Rep*, 17(3):319–333.

Wallace, I. M., O'Sullivan, O., Higgins, D. G., and Notredame, C. (2006). M-coffee: combining multiple sequence alignment methods with t-coffee. *Nucleic Acids Res*, 34(6):1692–1699.

Walle, I. V., Lasters, I., and Wyns, L. (2005). Sabmark–a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, 21(7):1267–1268.

Wallin, E. and von Heijne, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci*, 7(4):1029–1038.

Wang, G. and Dunbrack, R. L. (2004). Scoring profile-to-profile sequence alignments. *Protein Sci*, 13(6):1612–1626.

Waterman, M. S. and Eggert, M. (1987). A new algorithm for best subsequence alignments with application to trna-rrna comparisons. *J Mol Biol*, 197(4):723–728.

Whisstock, J. C. and Lesk, A. M. (2003). Prediction of protein function from protein sequence and structure. *Q Rev Biophys*, 36(3):307–340.

White, S. H. (2004). The progress of membrane protein structure determination. *Protein Sci*, 13(7):1948–1949.

Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. *Science*, 319(5862):473–476.

Wu, G., Chen, Y. G., Ozdamar, B., Gyuricza, C. A., Chong, P. A., Wrana, J. L., Massagué, J., and Shi, Y. (2000). Structural basis of smad2 recognition by the smad anchor for receptor activation. *Science*, 287(5450):92–97.

Xiong, J. (2006). *Essential Bioinformatics*, chapter Multiple sequence alignment, page 69. Cambridge University Press.

Xu, Z., Horwich, A. L., and Sigler, P. B. (1997). The crystal structure of the asymmetric groel-groes-(adp)7 chaperonin complex. *Nature*, 388(6644):741–750.

Yakymovych, I., Heldin, C.-H., and Souchelnytskyi, S. (2004). Smad2 phosphorylation by type i receptor: contribution of arginine 462 and cysteine 463 in the c terminus of smad2 for specificity. *J Biol Chem*, 279(34):35781–35787.

Ye, K., Lameijer, E.-W. M., Beukers, M. W., and Ijzerman, A. P. (2006). A two-entropies analysis to identify functional positions in the transmembrane region of class a g protein-coupled receptors. *Proteins*, 63(4):1018–1030.

Ye, Y. and Godzik, A. (2003). Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, 19 Suppl 2:ii246–ii255.

Yona, G. and Levitt, M. (2002). Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J Mol Biol*, 315(5):1257–1275.

Zachariah, M. A., Crooks, G. E., Holbrook, S. R., and Brenner, S. E. (2005). A generalized affine gap model significantly improves protein sequence alignment accuracy. *Proteins*, 58(2):329–338.

Zardoya, R. and Villalba, S. (2001). A phylogenetic framework for the aquaporin family in eukaryotes. *J Mol Evol*, 52(5):391–404.

Zhou, H. and Zhou, Y. (2005). Spem: improving multiple sequence alignment with sequence profiles and predicted secondary structures. *Bioinformatics*, 21(18):3615–3621.

Zuker, M. (1991). Suboptimal sequence alignment in molecular biology. alignment with error analysis. *J Mol Biol*, 221(2):403–420.

Samenvatting

Vergelijken van bouwstenen van het leven: sequentie alignment en evaluatie van voorspelde structurele en functionele eigenschappen.

Van DNA naar eiwit naar multiple sequence alignment.

Dit proefschrift beschrijft de resultaten van mijn vierjarig promotie-onderzoek naar geautomatiseerde methodes om eiwitten met elkaar te vergelijken en te analyseren. Dit roept allereerst de vraag op wat eiwitten nu precies zijn en welke rol vergelijkingstechnieken spelen bij de analyse ervan.

Aan de grondslag van onze eiwitproductie ligt het DNA, het materiaal waarin onze erfelijke eigenschappen besloten zitten. Het DNA bevat belangrijke stukken, genaamd genen, die 'coderen' voor één of meerdere eiwitten. Eiwitten zijn op hun beurt moleculen die een groot scala aan functies herbergen en die essentieel zijn voor het goed functioneren van cellen. Ze zijn opgebouwd uit blokjes die we aminozuren noemen. Op het DNA liggen de codes voor 20 verschillende aminozuren welke als het ware aan elkaar geregen worden tot eiwitketens. Een eiwit krijgt zijn uiteindelijke functionaliteit als de aan elkaar gekoppelde reeks aminozuren opvouwen tot een correcte tertiare structuur.

Zoals gezegd hebben organismen vele soorten eiwitten die verschillen in hun aminozuurketen. Deze verschillen zijn het resultaat van evolutionaire tijdsprocessen gedurende welke veranderingen in het DNA hebben geleid tot de huidige variëteit aan genetisch materiaal èn dus aan eiwitten. Stel dat we een bepaald eiwit X onder de loep nemen dat een belangrijke rol speelt in het stofwisselingsproces. Het is dan

niet alleen interessant om de sequentie van enkel het menselijke eiwit X te bekijken, maar ook om overeenkomstige eiwitten in andere soorten met eenzelfde functie erbij te betrekken. We gaan er hierbij vanuit dat al deze eiwitten afstammen van een oer-eiwit, welke door de tijd heen is geëvolueerd met als 'eindresultaat' de huidige familie van sequenties in de verschillende soorten. We noemen dit principe homologie: de onderzochte sequenties stammen af van eenzelfde voorouder.

Uiteindelijk heeft de originele oer-sequentie dusdanige veranderingen ondergaan dat er tussen de soorten verschillen zijn opgetreden in eiwit X, zowel wat betreft de lengte als de compositie van de aminozuurketen. Aan de andere kant zullen er ook bepaalde stukken sterk op elkaar lijken en zogenaamd geconserveerd zijn gebleven vanwege hun cruciale rol in het uitoefenen van de functie van eiwit X. Zodoende kunnen we door de eitwit sequenties van X in diverse soorten met elkaar te vergelijken erachter komen waar deze functioneel essentiële aminozuren nu precies zitten in het eiwit.

Een ander voorbeeld: stel dat men van eiwit X aanwijzingen heeft dat het in sommige afwijkende gevallen een rol speelt in de ontwikkeling van een bepaald type kanker. In een dergelijk geval zou het interessant zijn om de aminozuursequentie te achterhalen van zieke patiënten en deze te vergelijken met die van gezonde mensen. Wellicht kunnen we op deze manier achterhalen op welke plekken mutaties zich hebben voor gedaan die een aandeel zouden kunnen hebben in de vorming van de tumor.

Voordat we een analyse kunnen loslaten op onze eiwitsequenties en de stap kunnen gaan maken naar patroonherkenning, zullen we een programma nodig hebben dat de sequenties met elkaar kan vergelijken. Dit is een gecompliceerd proces aangezien er, zoals eerder gezegd, door de tijd heen verschillen zijn opgetreden zowel qua lengte als aminozuurcompositie. De uitdaging is dus om vragen te beantwoorden als: 'Welk stukje van het menselijk eiwit X correspondeert met een bepaalde regio van het homologe eiwit X van een muis?'. Voor het beantwoorden van zulke vragen is de kwaliteit van de vergelijkingsprogramma's cruciaal.

De aangewezen geautomatiseerde computertechniek die we gebruiken om meerdere sequenties met elkaar te vergelijken heet 'multiple sequence alignment'. Beknopt zou men kunnen zeggen dat deze methodes de verschillende eiwitten inlezen en vervolgens de corresponderende aminozuren zo proberen onder elkaar te zetten dat de het aantal overeenkomsten wordt gemaximaliseerd. Er wordt dus geschoven met de aminozuurblokjes en op plekken waar geen overeenkomsten worden gevonden zet de methode gaten die we 'gaps' noemen. Op deze manier krijgen alle input sequenties dezelfde lengte en kan het eindresultaat, het multiple alignment, worden gezien als een tweedimensionale metrix met lengte L (totaal aantal aminozuren en gaps) en breedte N (het aantal sequenties), zoals is weergegeven in het voorbeeld van Table 9.1. Een uitgebreid overzicht van multiple sequence alignment methodes en hun basisprincipes staan beschreven in Hoofdstuk 2.

mens	**M**	P	I	**R**	**N**	I	A	**G**	**R**	**P**	D	E	**A**	**T**	R	L
hond	**M**	S	H	**R**	**N**	H	A	**G**	**R**	**P**	G	L	**A**	**T**	V	-
muis	**M**	-	L	**R**	**N**	I	-	**G**	**R**	**P**	D	E	**A**	**T**	K	L
kat	**M**	P	I	**R**	**N**	-	-	**G**	**R**	**P**	D	E	**A**	**T**	A	-

Table 9.1: Voorbeeld van een multiple sequence alignment van eiwit X in vier verschillende soorten. De vetgedrukte letters wijzen op volledig geconserveerde aminozuren die waarschijnlijk essentieel zijn voor het functioneren van het eiwit.

De structuur van een eiwit wordt beter geconserveerd dan de aminozuursequentie.

In de vorige paragraaf werd kort aangestipt dat eiwitten pas functioneel worden in de cel op het moment dat hun primaire aminozuurketen zich opvouwt tot een tertiare structuur. Als we kijken naar conservering van deze twee niveaus, zien we dat structuren door de tijd heen meestal beter worden geconserveerd dan de sequenties. Met andere woorden, twee eiwitten kunnen grote verschillen vertonen tussen hun aminozuren maar toch nog opvouwen tot een nagenoeg gelijke driedimensionale structuur. Zo zijn er vele gevallen bekend van eiwitten met een bijna identieke structuur hoewel meer dan 70% van hun aminozuren met elkaar verschillen. Dit roept de vraag op of het daarom niet beter is om structurele informatie van eiwitten te gebruiken bij het vergelijken van hun sequenties. In principe is dat een goed idee, ware het niet dat het bepalen of 'oplossen' van een eiwitstructuur een ingewikkeld en tijdrovend proces is in vergelijking tot het bepalen van een aminozuursequentie. Ter vergelijking: op het moment zijn er van miljoenen eiwitten de sequenties bekend, terwijl het aantal opgeloste tertiare structuren wereldwijd rond de 60.000 ligt.

Uiteraard is er in de afgelopen decennia veel onderzoek geweest naar het vouwingsproces en heeft men geprobeerd te begrijpen op welke wijze de aminozuren nu precies een structuur vormen. Dit heeft geresulteerd in vele structuurvoorspellingsmethoden. Echter is het tot op heden in vrijwel alle gevallen nog niet mogelijk om van enkel de aminozuursequentie te voorspellen hoe een eiwit zich precies zal opvouwen. Desalniettemin is het een stuk eenvoudiger gebleken om een intermediar stadium te voorspellen: de secundaire structuur.

Tijdens het vouwingsproces organiseert de eiwitketen zich namelijk aanvankelijk tot drie verschillende typen ruimtelijke vormen die tezamen de secundaire structuur bepalen: de alfa-spiraal (helix), de bèta-structuur (strand of sheet) en een restgroep die coil wordt genoemd. De eerste twee structuren worden door waterstofbruggen in hun ruimtelijke vorm gehouden. Ook bij deze structuurvorm gaat het evolutionaire conservatieprincipe op: een secundaire structuur is door de tijd heen beter geconserveerd dan een primaire eiwitsequentie. Dit betekent dat, ondanks het feit dat we vaak geen betrouwbare informatie over de tertiare eiwitstructuur hebben, we toch goed de secundaire structuur kunnen voorspellen op basis van de primaire aminozuurketen.

Het multiple alignment programma PRALINE combineert primaire sequentie data met voorspelde secundaire structuurinformatie.

Een groot deel van dit proefschrift is gewijd aan het de ontwikkeling van het multiple sequence alignment programma PRALINE. Het nieuwe element in dit programma is het gebruik van voorspelde secundaire structuur bij het onder elkaar zetten van de eitwitsequenties. Samengevat wordt er van elke ingevoerde eiwitsequentie de secundaire structuur voorspeld met behulp van een externe tool. De secundaire structuur, geclassificeert als alfa-spiraal (helix), bèta-structuur of coil, wordt samengevoegd met de primaire sequentie informatie. Uiteindelijk is dan voor elke aminozuurletter een voorspelling voor handen tot welke secundaire structuur catergorie deze behoort. PRALINE gebruikt hierna een op maat gesneden scoringsschema om zo goed mogelijk de letters en de bijbehorende secundaire structuren op elkaar te passen. Bij het vergelijken van de aminozuren wordt zodoende niet alleen naar de letters *an sich* gekeken, maar ook 'geluisterd' naar de secondaire strucuur die meer geconserveerde en sterkere overeenkomsten vertoont tussen de sequenties.

Daarnaast is er binnen PRALINE nog een extra protocol ingebouwd dat specifiek is bedoeld voor het alignen van een speciale groep eiwitten: de transmembraan eiwitten. Deze groep kenmerkt zich door de aanwezigheid van een extra type secundaire structuur, naast de boven beschreven types. Zoals de naam 'transmembraan' suggereert, hebben deze eiwitten met elkaar gemeen dat een gedeelte van hun eiwit door het cellulair membraan heen gaat. Transmembraan-eiwitten zijn over het algemeen belangrijke receptoren en spelen vaak een rol in de ontwikkeling van kanker. In het merendeel van de gevallen organiseert het transmembraan gedeelte zich in een karakteristieke alfa-helix. De locatie van deze gedeeltes kan tegenwoordig redelijk nauwkeurig worden voorspeld op basis van de primaire sequentie. PRALINE maakt hievan gebruik op een analoge manier als in het geval van 'gewone' secundaire structuur voorspelling. Echter nu wordt het alignment van de primaire aminozuurketens voornamelijk geleid door de informatie verkregen uit de voorspellingen van de transmembraan segmenten.

Een ander voordeel van deze aanpak betreft het feit dat er tijdens het alignen van de afzonderlijke aminozuren nu ook impliciet rekening wordt gehouden met hun 'buren'. Secundaire structuur voorspellingstechnieken gebruiken namelijk meerdere aminozuren tegelijk om tot de voorspelling van één positie te komen. Dit is ook logisch want bijvoorbeeld een alfa-spiraal beslaat ook meerdere posities.

In de Hoofdstukken 3 en 4 staan de werkwijze van PRALINE en de behaalde resultaten beschreven. Hierin komt de toegevoegde waarde van de geïntegreerde secundaire en transmembraan structuur voorspellingen duidelijk naar voren. Ook uit de vergelijking met andere veel gebruikte multiple sequence alignment methoden kan worden geconcludeerd dat de combinatie van primaire en secundaire sequentie informatie een grote toegevoegde waarde heeft voor de kwaliteit van het uiteindelijke alignment. De methode is beschikbaar die online gebruikt worden op webadres: `www.ibi.vu.nl/programs/pralinewww`.

Analyse van multiple sequence alignments: voorspelling van subgroep-specifieke aminozuren met behulp van de Sequence Harmony methode.

Een ander belangrijk deel van dit proefschrift is gewijd aan de analyse van multiple alignments. Er is hierbij meer gericht gekeken naar aminozuren die karakteristiek zijn voor bepaalde eiwit-subgroepen. In het algemeen worden eiwitfamilies onderverdeeld in subgroepen op basis van verschillen in functie. Als we (opnieuw) een voorbeeld-eiwit X erbij nemen dat een aandeel heeft in stofwisselingsprocessen, dan zou het kunnen zijn dat er een subgroep X-1 bestaat die zorgt voor een wat snellere stofwisseling in vergelijking tot eiwitten in subgroep X-2. Of bijvoorbeeld een geval waarbij binnen een bepaalde groep mensen eiwit X zich anders manifesteert dan bij gezonde mensen. In deze gevallen is het zeer interessant om te kijken of we op het aminozuurniveau een verklaring kunnen geven voor deze verschillende functies binnen dezelfde familie X. We zijn hierbij dus op zoek naar functionele plekken binnen het eiwit die verantwoordelijk zijn voor het verschil tussen de subgroepen. We noemen deze 'specificiteit-bepalende aminozuren'.

Deze specificiteit kan het beste worden bepaald wanneer we per subgroep meerdere sequenties tot onze beschikking hebben. Van alle aangeleverde sequenties wordt dan eerst een alignment gemaakt, bijvoorbeeld met behulp van PRALINE, en vervolgens worden de sequenties ingedeeld per subgroep (we gaan er hier van uit dat de classificatie voor elke sequentie bekend is). Reeds geïntroduceerde methodes uit vorige studies, eveneens gemaakt met het doel om van alignments 'specificiteit-bepalende aminozuren' te vinden, gaven hierbij een grote prioriteit aan 'conservering'. Dat wil zeggen dat deze methodes voornamelijk op zoek gingen naar aminozuren die welliswaar verschillen toonden tussen de subgroepen, maar wél geconserveerd waren binnen de subgroepen. Echter in de praktijk blijkt deze conservering niet altijd bepalend te zijn: er zijn ook veel plekken die zowel verschillen tussen als binnen de subgroepen vertonen. Het Sequence Harmony algoritme baseert zich dan ook niet op conserveringsprincipes, maar stelt als enige vereiste dat de aminozuren tússen de subgroepen van elkaar moeten verschillen. Een nieuw ontwikkelde formule geeft een score aan alle alignment posities, waarbij een minimale waarde van '0' wordt toegekend aan specificiteit-bepalende aminozuren en een maximale waarde van '1' aan 'onspecifieke' posities.

Hiernaast houdt de Sequence Harmony methode ook rekening met het feit dat specifieke functies over het algemeen worden uitgevoerd door meerdere aminozuren tegelijk. De sterkte van het signaal wordt derhalve mede bepaald door het aantal specificiteit-bepalende aminozuren op een gegeven segment van het eiwit. Het programma is getest op een aantal experimenteel gevalideerde testcases en heeft aangetoond accuratere voorspellingen te geven dan eerder ontwikkelde methodes op dit gebied. We kunnen concluderen dat de kracht van de nieuw ontwikkelde methode ligt in een combinatie van 1) de nadruk die wordt gelegd op daadwerkelijke verschillen tussen eiwit-groepen in plaats van alleen te kijken naar geconserveerde verschillen en 2) het feit dat er een hogere prioriteit wordt gegeven aan het specificiteits-signaal van een groep posities dan aan het signaal van 'alleenstaande' posities. De resultaten van

dit onderzoek staan beschreven in de Hoofdstukken 5 en 6.

Er is ook een interactieve website gemaakt met behulp waarvan eiwitfamilies kunnen worden geanalyseerd op subgroep-specificiteit: `www.ibi.vu.nl/programs/seqharmwww`.

Wat is de betekenis van alignments? Dynamische eiwitten *versus* statische alignments.

Het laatste gedeelte van dit proefschrift betreft een studie naar de betekenis en interpretatie van alignments. We kijken hierbij vooral naar de relatie tussen statische aminozuur-alignments en de beweeglijkheid van eiwitten.

Eerder is beschreven hoe de primaire aminozuurketens zich stapsgewijs opvouwen, via een secundaire structuur, tot een tertiair niveau waarbij eiwitten hun functionaliteit krijgen. Hieraan moet worden toegevoegd dat eiwitten niet statische, maar juist zeer beweeglijke moleculen zijn. Deze flexibiliteit ik ook van groot belang om functies goed uit te kunnen oefenen. Zo moeten eiwitten op sommige momenten bijvoorbeeld aan elkaar of aan metalen binden om actief te worden, en op andere momenten juist een meer passieve rol aannemen. Als gevolg hiervan kan men stellen dat eiwitten geen 'unieke' tertiare structuur hebben en dat de kristallisatie van structuren zich slechts beperkt tot het maken van een momentopname van de beweeglijke eiwitten.

Aan de andere kant kunnen we nog steeds veel belangrijke informatie halen uit deze momentopnames en zijn deze waardevol voor talloze doeleinden. Bijvoorbeeld multiple sequence alignment methodes maken hier vaak gebruik van om de methodes mee te testen. Dit vloeit voort uit het eerder genoemde principe dat dat het conservatieniveau van tertiare structuren hoger stelt dan dat van de onopgevouwen primaire aminozuurketens. Zo zijn er vele referentie-alignments van eiwitten beschikbaar die gemaakt zijn met behulp van tertiare structuur informatie. Het is namelijk makkelijker om de overeenkomsten tussen eiwitstructuren te detecteren en vervolgens terug te mappen op het sequentieniveau.

Bij structurele vergelijkingen wordt geprobeerd de eiwitten zo goed mogelijk op elkaar te passen. Echter, het feit dat eiwitstructuren dynamisch zijn zorgt voor een dilemma, aangezien de beste oplossing afhankelijk is van het tijdsmoment waarop het eiwit is gekristalliseerd. Op het moment dat we nu gebruik maken van structurele vergelijkingen om sequentie alignment routines te testen, betekent dit dat hier ook fouten zullen worden gemaakt. Een multiple alignment geeft namelijk ook niet meer weer dan een statische vergelijking tussen eiwit sequenties.

In Hoofdstuk 7 en 8 worden deze problemen onderzocht en testen we invloed van eiwitbewegingen op van de structuur afgeleide sequentie-alignments. We zien dat zelfs lichte veranderingen in de eiwitstructuur kunnen leiden tot grote variaties in de alignments. Bij de functionele interpretatie van alignments moet dus rekening worden gehouden met het feit dat deze geen eenduidig antwoord geven op de werkelijke situatie. Er worden in deze hoofdstukken ook nieuwe oplossingen aangedragen om multiple sequence alignment methodes te testen rekening houdend met eiwitflexibiliteit.

Acknowledgements

Many people have professionally and non-professionally contributed to the achievement of this very special and gratifying moment of my life.

My first thoughts go to Jaap who gave me the opportunity to perform the project. I have always felt very privileged to be supervised and guided such a kind, interesting and, above all, interested person. Jaap has not only given me great advice on all kinds of matters, but he has also helped me to become a complete and self-confident bioinformatician. Furthermore I am deeply impressed by the way in which he has created an extraordinary team of (young) talented people. In such a positive atmosphere it is impossible not to enjoy your time and I am convinced that this was a big help in successfully completing my PhD.

When I started this journey, my fellow passenger from the very first day was Anton Feenstra. He started at the same time as me at the IBIVU as a Post-doctoral researcher and the first two years we had a great time in room P4.40. Unforgettable moments like drawing the 'Sequence Harmony' formula on the white board and listening to songs of the huge music database one will hardly ever forget. Obviously I am also very grateful for the many discussions and enormous help you have given me during the past years and you certainly can claim a great share in this thesis work.

Also I would like to thank my extraordinary colleagues in the IBIVU who have contributed hugely to a wonderful PhD period. Thomas, Hannes, Bernd and Bart, I am very grateful to have met such a good friends. I think we have a lot of great and crazy moments to remember, and I am sure these will also continue after P1.38! I thank my two paranymphs, Daan and Thomas for always standing right (and left) next to me, even in the moments of my defense.

I would like to express my gratitude to all members of the committee, prof. Des Higgins, prof. William Taylor, prof. Roland Siezen, prof. Jack Leunissen, prof. Bas Teusink and dr. Sanne Abeln. Furthermore I would like to acknowledge the Netherlands Bioinformatics Centre (NBIC) for funding this project.

Mijn laatste woorden van dank gaan uit naar mijn biologische en niet-biologische ouders. Hans en Petra, zonder jullie onvoorwaardelijke steun en liefde, van onze eerste stappen in Amsterdam tot onze huidige passen, had deze PhD zeker niet de blijdschap gekregen die het nu heeft. A Carlo ed Anna che ci hanno sempre seguiti da lontano, stando dietro la libertà di ogni nostra scelta guardando costantemente alla nostra vita con uno sguardo orgoglioso.

Vorrei dedicare però questa tesi a Valeria che ha avuto il grande coraggio di partire insieme a me da Milano. Come sempre hai dimostrato di avere tanto carattere nell'affrontare le scommesse della vita. Mi commuovo se penso a tutti i momenti in cui mi sei stata vicina, sempre con tanto amore, anche nei giorni difficili di solitudine. Anni fa abbiamo iniziato a sognare insieme e ci siamo posti diversi obbiettivi (forse anche troppi). La nostra unione, che poi è la nostra forza, ci ha portati a concretizzarne molti ed ogni giorno di più capisco che non potrei vivere neanche un attimo senza di te!

Publications

Pirovano, W., Simossis, V.A., and Heringa, J., Secondary structure-guided multiple sequence alignment, *submitted*.

Pirovano, W., and Heringa, J., Protein secondary structure prediction, *Methods Mol. Biol.*, in press.

Pirovano, W., van der Reijden, A., Feenstra, K.A., and Heringa, J. (2009), Structure and function analysis of flexible alignment regions in proteins, *BMC Bioinformatics*, 10(Suppl 13):P6.

van Houte, B.P.P., Binsl, T.W., Hettling, H., Pirovano, W., and Heringa, J. (2009), CGHnormaliter: an iterative strategy to enhance normalization of array-CGH data with imbalanced abberations, *BMC Genomics*, 10:401.

Pirovano, W., Feenstra, K.A., and Heringa, J. (2008), The meaning of alignment: lessons from structural diversity, *BMC Bioinformatics*, 9:556.

Pirovano, W., and Heringa, J. (2008), Multiple Sequence Alignment, *Methods Mol. Biol.*, 452:143–161.

Pirovano, W., Feenstra, K.A., and Heringa, J. (2008), PRALINE$^{\mathrm{TM}}$: a strategy for improved multiple alignment of transmembrane proteins, *Bioinformatics*, 24(4):492–497.

Horner, D.S., Pirovano, W., and Pesole, G. (2008), Correlated substitution analysis and the prediction of amino acid structural contacts, *Brief. Bioinform.*, 9(1):46–56.

Heringa, J., and Pirovano, W. (2007), Sequence similarity searches, *Bioinformatics, Method Express Series*, Dear, P. ed., Scion Publishing Ltd, Oxfordshire, UK, 39–67.

Feenstra, K.A., Pirovano, W., Krab, K., and Heringa, J. (2007), Sequence harmony: detecting functional specificity from alignments, *Nucleic Acids Res.*, 35(Web Server issue):W495–W498.

Marchiori, E.*, Pirovano, W., Heringa, J., and Feenstra, K.A.* (2006), A feature selection algorithm for detecting subtype specific functional sites from protein sequences for Smad receptor binding, *The Fifth International Conference on Machine Learning and Applications (ICMLA'06), IEEE*, 168–173. (* equal contribution)

Pirovano, W.*, Feenstra, K.A.*, and Heringa, J. (2006), Sequence comparison by sequence harmony identifies subtype-specific functional sites, *Nucleic Acids Res.*, 34(22):6540–6548. (* shared first authors)

Curriculum Vitae

Walter Pirovano was born on April 5, 1980. After completing his years at the Coornhert Gymnasium in Gouda, he attended the University of Milan to study biology. During his specialization in molecular biology he joined the bioinformatics lab of Prof. Graziano Pesole for a thesis project. Here he has been working on correlated amino acid substitution analysis under the daily supervision of Dr. David Horner. In 2005 he obtained his Master's degree and moved to the VU University Amsterdam to start his PhD research in the group of Prof.dr. Jaap Heringa. Under his supervision and the co-supervision of Dr.ir. Anton Feenstra, Walter performed studies in the field of sequence analysis for four years with a primary focus on multiple sequence alignment. The work presented in this thesis includes the main research results obtained over that period.

Currently, Walter is working as a bioinformatician in the genome analysis section of BaseClear in Leiden. The main focus here is management and analysis of next-generation sequencing data.